SCHOOL OF ECONOMICS AND MANAGEMENT

LUND UNIVERSITY

Leveraging LlaMA 2 for sentiment analysis

A study on OMXS30 stocks

Johan Darell
jo4531da-s@student.lu.se

Vivea Upadhyaya
vi8300up-s@student.lu.se

NEKH02
Bachelor's Thesis (15 credits ECTS)
20th January 2024
Supervisor: Andreas Johansson, andreas.johansson@nek.lu.se

Abstract

This thesis investigates the application of sentiment analysis in predicting stock returns for the companies listed in the OMXS30 index. The recent development of large language models (LLMs), like ChatGPT, has substantially advanced the field of sentiment analysis. This thesis utilizes Meta's LlaMA 2 LLM for sentiment analysis, while a random forest model is employed to predict monthly stock returns for the subsequent month. These predictions enable the creation of a dynamic long-short portfolio, with its returns evaluated against the OMXS30 index using the capital asset pricing model, the Fama–French three-factor model, and the Carhart four-factor model. The results demonstrate that it is possible to generate excess returns by leveraging LlaMA 2, with sentiment ingrained as a key risk factor. The thesis contributes to the existing literature by examining the Swedish stock market as well as using the open-source LLM LlaMA 2.

Acknowledgements

We wish to express our most sincere appreciation and gratitude to our supervisor, Andreas Johansson, for his invaluable guidance and patience throughout the development of this thesis. We are also grateful to Manu Upadhyaya for his inputs and insights, and to our classmates for their engaging discussions and support.

Johan Darell, Vivea Upadhyaya
Lund, January 2024

Abbreviations

Abbreviation	Expansion
AGI	Artificial general intelligence
AI	Artificial intelligence
BERT	Bidirectional encoder representations from transformers
BLUE	Best linear unbiased estimator
C4F	Carhart's four-factor model
CAPM	Capital asset pricing model
CV	Cross-validation
EMH	Efficient market hypothesis
FF3F	Fama–French three-factor model
HML FF	High minus low based on Fama–French
LlaMA	Large language model Meta AI
LLM	Large language model
LSTM	Long short-term memory
MAE	Mean absolute error
MKT	Market risk
ML	Machine learning
MSE	Mean square error
NLP	Natural language processing
OLS	Ordinary least squares
SA	Sentiment analysis
SMB	Small minus big
UMD	Up minus down

Contents

1 Introduction

The endeavor to predict stock returns essentially aims to "beat the market" or outperform the broader market. Yet, the efficient market hypothesis (EMH) contends that achieving persistent excess returns through informational advantages is practically impossible since stock prices are thought to adequately represent all available information (Fama, 1970). However, recent advances in natural language processing (NLP), particularly with large language models (LLMs) such as Meta's LlaMA 2, has increased the use and prominence of these technologies in exploring the extent to which stock prices truly encompass all available information.

The modern financial landscape is rapidly evolving, with sentiment playing an important part in determining market dynamics. Recent studies including Fatouros et al. (2023), Kearney and S. Liu (2014), and Breitung, Kruthof and Müller (2023) have underscored this influence. This thesis places a strong emphasis on sentiment analysis, particularly through the use of the advanced LlaMA 2 language model, mirroring approaches seen in Pavlyshenko (2023)'s research. It aims to provide a novel perspective on understanding and leveraging these market dynamics. By investigating the predictive power of sentiment, this thesis contributes to the growing body of knowledge on the impact of sentiment as a risk factor in stock price performance.

The core research questions addressed in this thesis are twofold. First, does investor sentiment, as determined through sentiment analysis of company-specific news article headlines from the year 2019 to 2023, accurately predict stock returns for companies listed on the OMXS30 index? Second, can a constructed portfolio based on these predicted excess returns, evaluated with the capital asset pricing model (CAPM), the Fama–French three-factor model (FF3F), and the Carhart four-factor model (C4F), outperform the market? This follows with a discussion on whether sentiment is a missing risk-factor in the mentioned asset pricing models.

The methodology includes collecting and preprocessing 60,871 news headlines related to specific OMXS30 index tickers. The 2019-2023 analysis period was chosen for its relevance to current financial landscapes and significant macroeconomic events. LlaMA 2 was used for sentiment annotation, with headlines classified as positive, negative, or neutral. A random forest model is used to predict stock returns, using monthly sentiment and daily returns data. To maintain chronological order and real-life prediction scenarios, the data split for training, testing and validation sets uses a time series method. The predicted returns are then used to create a dynamic portfolio, finally evaluated against the CAPM, FF3F and C4F.

This research enhances existing knowledge by focusing on the Swedish stock market and using the open-source LLM LlaMA 2 for sentiment analysis. It provides insights into the impact of investor sentiment on OMXS30, integrating sentiment data into financial analysis and challenging conventional ideas of market efficiency and predictability. The findings suggest that investor sentiment, a risk factor overlooked by traditional models like CAPM, FF3F, and C4F, is crucial for understanding market dynamics and developing effective investment strategies. Moreover, the thesis's emphasis on the OMXS30 index adds valuable insights into a relatively underexplored market in sentiment analysis research.

The rest of this paper is organized as follows. Chapter 2 delves into the theoretical background and related literature, encompassing topics like the EMH, various methods of stock prediction, sentiment analysis, machine learning with a particular focus on the random forest model, and the CAPM along with its multi-factor extensions. Chapter 3 details the methodology employed in the thesis, including the processes of sentiment data collection and preprocessing, analysis of stock price data, the methodology behind sentiment analysis, the development of the prediction model using random forests, and the construction and evaluation of portfolios based on sentiment analysis with the CAPM, FF3F, and C4F models. In Chapter 4, the results are presented, focusing on the predictions generated by the model and the performance of the constructed portfolio. Chapter 5 offers a discussion of the findings, exploring their implications, limitations, significance, and possibilities for future research. Finally, Chapter 6 concludes the thesis, summarizing the main discoveries and contributions of the thesis.

2 Theory and related literature

The efficient market hypothesis posits that stock prices fully reflect all available information, a theory complemented by behavioral finance, which emphasizes the role of sentiment in financial decisions. The chapter delves into sentiment analysis and its intersection with behavioral finance, emphasizing its importance in predicting stock returns. Furthermore, traditional and modern stock prediction methods are investigated, including the use of machine learning. Lastly, the capital asset pricing model (CAPM), the Fama–French three-factor model (FF3F), and the Carhart four-factor model (C4F) are covered.

2.1 The efficient market hypothesis

The efficient market hypothesis (EMH) assumes that all available information is fully reflected in stock prices, indicating informational efficiency in the market. The EMH is segmented into three forms; weak, semi-strong, and strong. The weak form asserts that all historical information is factored into stock returns. The semi-strong form includes both historical and all publicly available information. The strong form extends to insider information as well. The random walk hypothesis, which suggests that stock return fluctuations in an efficient market are random and unexpected, is a key component of the efficient market hypothesis (Fama, 1970). Fama's hypothesis implies that surpassing market returns, as represented by a general economic benchmark, on a risk-adjusted basis through stock selection or market timing is not possible (Malkiel, 2003). This perspective, however, does not entirely preclude the possibility of predicting returns, but rather emphasizes the difficulty in predicting risk-adjusted returns.

Challenges to the EMH, as noted by Malkiel (2003) and Shiller (2003), include phenomena such as stock market bubbles and crashes. Additionally, Barberis and Thaler (2003) highlights the impact of cognitive biases and emotional factors on financial decisions. Despite these challenges, Malkiel (2003) acknowledges that, even in the face of irregularities and market imperfections like the 1999 'bubbles,' the stock market generally exhibits remarkable efficiency in utilizing information. This efficiency implies that any anomalies or predictable patterns are unlikely to remain consistent or significant enough to yield consistently exceptional returns.

Behavioral finance, integrating psychological insights into economic and financial theories, provides a nuanced understanding of financial markets, as indicated by Shiller (2003) and Bird, Du and Willett (2017). Shiller (2000) in "Irrational Exuberance" argues that investor sentiment, often influenced by news media, can significantly impact stock returns by shaping future price expectations. This aligns with behavioral finance, suggesting that sentiment analysis reveals the emotional and psychological factors affecting market dynamics. Empirical studies that show the impacts of news events on stock return movements, like those in Bollen, Mao and Zeng (2011), Bondt and Thaler (1985) and Ding et al. (2014), lend support to this idea.

Traditional EMH models do not explicitly account for sentiment as a risk factor. Nonetheless, studies like Boido and Fasano (2014) demonstrate sentiment's substantial effect on investors' risk perceptions and decisions, prompting a potential re-evaluation of how risk premiums are assessed in the market. Consequently, this thesis acknowledges the principles of EMH in recognizing market efficiency but also explores the possibility that investor sentiment may be an overlooked risk factor in major asset pricing models, including CAPM, FF4F, and C4F. This investigation aims to bridge traditional EMH perspectives with new insights into the impact of psychological factors in financial markets. The role of investor sentiment thus becomes an interesting point of contrast when viewed through the lens of the EMH.

2.2 Stock prediction

Stock prediction is widely utilized by both researchers and investors as a key tool for achieving financial gains in trading (Ferreira, Gandomi and Cardoso, 2021). The predictability of a stock's future return is a complex, controversial, and extensively researched topic, drawing interest from disciplines such as finance, mathematics, and computer science (Madge and Bhatt, n.d.), where the controversy stems from the assumptions of the EMH, and also negative empirical evidence that give rise to a debate about the performance of predictions (Park and Irwin, 2007).

Fundamental analysis and technical analysis are two traditional methods that have been foundation pillars for stock predictions, whereas modern methods for stock prediction use artificial intelligence (AI) and sentiment analysis (SA) (Ferreira, Gandomi and Cardoso, 2021; Rouf et al., 2021). Fundamental analysis entails studying the intrinsic value of a security through financial and macroeconomic factors, while technical analysis only studies historical prices and volumes predict future price movements. Sentiment analysis involves using stock-related information such as news articles, financial statements, and commentaries to identify investor sentiment towards a stock or the stock market (Huang, Capretz and Ho, 2021). With the integration of these analyses, AI and its subset machine learning (ML), can be applied to enhance prediction accuracy (Ferreira, Gandomi and Cardoso, 2021; Rouf et al., 2021).

Due to the erratic nature of the stock market and inherent features like volatility, noise, and irregularities, predicting stock returns is a challenge (Gandhmal and Kumar, 2019; Rouf et al., 2021). It's crucial to understand that the stock market is complex and affected by a wide range of interconnected factors, the movements in a stock could be the result of many possible factors; the company's performance, the social and political situation, the local and global economic situation and market sentiment (Branch, 1976; Eleswarapu and Reinganum, 2004; Inthachot, Boonjing and Intakosum, 2016). The accuracy of stock prediction relies on several key factors such as the quality of the data, features or variables with a strong correlation with stock return movements and models that are sophisticated enough to capture complexities without overfitting the data (Kuhn and K. Johnson, 2013; Mishev et al., 2020; Shmueli, 2010; Ticknor, 2013).

In Bondt and Thaler (1985) and Eleswarapu and Reinganum (2004), they show that the influence of investor sentiment on stock returns is significant. Using sentiment analysis in a stock market context, research works such as Bollen, Mao and Zeng (2011) and Shapiro, Sudhof and Wilson (2020) have found a positive correlation between investor sentiment and stock returns. Additionally, these studies offer empirical evidence that supports stock prediction's feasibility using sentiment analysis.

2.3 Sentiment analysis

SA is an ongoing field of research that involves computationally decoding and interpreting people's opinions, sentiments, attitudes, and emotions as expressed in textual formats, which is done using tools such as natural language processing (NLP) and data mining. Usually, the texts are classified as being either positive, negative, or neutral (Araci, 2019). The analysis encompasses a variety of text formats, such as product reviews, news articles, and social media posts like tweets, and the core of SA lies in identifying and categorizing the underlying sentiment in these texts. SA is becoming more and more relevant and it is crucial for businesses to understand consumer perceptions and market trends, but it is also essential for broader societal contexts, like observing social dynamics or determining public opinion (Godsay, 2015; B. Liu, 2017; Medhat, Hassan and Korashy, 2014). As we transition from EMH's theoretical foundations, SA offers a unique viewpoint for comprehending the integration of information into market behavior (Kearney and S. Liu, 2014).

The application for sentiment analysis is wide, ranging from analyzing customer satisfaction to predicting stock returns or for the scope of this thesis: predicting stock returns (Godsay, 2015). Prior to determining the sentiment of a text, two key objectives need to be taken into account: domain-specificity and complexity. Complexity involves the different ways that a text can be written, including negations where a text might state "This is not good", as well as in more compositional phrases like "I wish I could have said I liked it", whereas domain-specificity involves the context since the same words can have different meanings depending on the context.

Sentiment analysis in finance is conducted using two main methods. The first method, lexical-based analysis, employs pre-defined dictionaries of words assigned with sentiments. However, Shapiro, Sudhof and Wilson (2020) has discussed various limitations to this method, as certain words may have different connotations in different contexts, potentially leading to inaccurate sentiment assessments. The second method makes use of ML techniques, which are classified as supervised or unsupervised learning. Supervised learning is based on pre-annotated datasets such as the FinancialPhraseBank (Malo et al., 2014), which contains around 5,000 sentences with sentiments assigned by financial experts and is widely used in academia (Araci, 2019; Colacicchi, 2022; Soong and Tan, 2021). Unsupervised learning, on the other hand, is used when there is no pre-existing annotated corpus, necessitating alternative techniques for sentiment determination (Bousquet, von Luxburg and Rätsch, 2004). Models built using ML, such as Reuters NewsScope Sentiment Engine, typically achieve higher accuracy compared to lexical-based methods as well as exceeding human analysts' assessments (Kearney and S. Liu, 2014)

2.3.1 Natural language processing

Natural language processing (NLP) is a process enabling computers to understand, interpret, and manipulate human language in a way that is both meaningful and useful. It was initially developed in the 1950s, but over time NLP borrowed concepts from various fields of science, mostly from statistics which began its application in the 1980s with the increasing prominence of machine learning. This era was characterized by the use of large annotated text corpora to train machine learning algorithms. Traditional and modern models enable systems to process human language in the form of text or voice data and to "understand" its full meaning, complete with the speaker's or writer's intentions and sentiment. NLP is a pivotal component in numerous applications ranging from automated chatbots to sophisticated data analysis in various domains, including finance (Xing, Cambria and Welsch, 2018b).

The utilization of textual data for modeling financial market dynamics has been a longstanding tradition in trading. This approach took a significant turn in the 1980s with the growing volume of financial reports, press releases, and news articles, leading to a paradigm shift towards automated analysis for maintaining a competitive edge in business. The correlation between Dow Jones daily returns and its historical data, for instance, began to diminish from the 1990s, prompting a transition towards more sophisticated mining models and linguistic techniques for financial forecasting (Xing, Cambria and Welsch, 2018b).

2.3.2 Financial sentiment analysis

Financial sentiment analysis, initially delineated in Stone et al. (1966), involves extracting and quantifying emotional tones from textual financial data such as corporate disclosures, media articles, and social media posts. This process has become increasingly relevant in behavioral finance, aiding in the interpretation of how sentiment impacts individual decision-makers, institutions, and markets (Kearney and S. Liu, 2014).

One sophisticated ML model that was developed by Google is BERT (Devlin et al., 2019) and thereafter its finance-specific adaptation FinBERT, developed by Araci (2019), marked a significant shift in how textual data is processed and interpreted in financial markets. BERT represented a pivotal advancement in NLP, analyzing the context of words in a sentence bidirectionally to better grasp the nuances of human language, meaning it looks at the surrounding context of words in sentences, not just the words alone, to better understand their meaning. FinBERT tailors this approach specifically for financial texts (Araci, 2019).

FinBERT's methodology encompasses several stages, including the use of long short-term memory (LSTM) neural networks. A neural network is an architecture that mimics the neural connections of the human brain. Because of their capacity to store information over extended periods and handle long-term dependencies, these networks excel in processing sequential and time-series data, making them ideal for tasks such as SA (Istiake Sunny, Maswood and Alharbi, 2020; Wang et al., 2016). The architecture of BERT, and its derivatives like FinBERT, leverage transformer encoders, a design that helps them understand language more effectively by focusing on the relationships between all words in a sentence (Araci, 2019). Furthermore, training strategies to prevent catastrophic inference, also known as catastrophic forgetting, which is the tendency of neural networks to forget previously learned information upon learning new information, are employed to enhance the performance of BERT and similar models (Devlin et al., 2019). These techniques ensure that the models retain the fundamental language information while adapting to the specific requirements of financial sentiment analysis. However, FinBERT can sometimes struggle with distinguishing between positive, neutral, and negative sentiments, especially in cases where financial statements are ambiguous or lack clear directional indicators (Araci, 2019). Despite this, FinBERT has long been the exemplar for sentiment analysis in the financial sector, being utilized in many research papers (Desola, Hanna and Nonis, 2019; J. Kim, H.-S. Kim and Choi, 2023).

In Fatouros et al. (2023) they delve into ChatGPT's capabilities in financial SA, particularly within the foreign exchange market. ChatGPT, an LLM by OpenAI, was evaluated without any fine-tuning or training for financial contexts, demonstrating its versatility and contextual comprehension and thus showcasing the potential of LLMs. The study found that ChatGPT outperformed FinBERT by approximately 35% in sentiment classification and exhibited a 36% higher correlation with market returns. Moreover, Fatouros et al. (2023) emphasizes the need for effective prompt design in zero-shot learning, indicating an absence of prior fine-tuning, and recognizes the difficulties in distinguishing complex sentiment tones. Overall, ChatGPT appears to be a promising tool for financial sentiment analysis, with an opportunity for further improvement and applications in this field.

2.3.2.1 Large language models

Language models (LMs) have been fundamental in natural language processing (NLP) tasks, particularly in predicting the next word in a sequence of text. Early models, such as statistical language models, utilized probabilistic approaches like n-grams to predict the next word based on the preceding $n - 1$ words. This evolved into more advanced deep learning architectures, including long short-term memory (LSTM) networks. A significant advancement occurred in 2017 with Vaswani et al. (2017) introducing the transformer architecture, catalyzing the development of large language models (LLMs) such as OpenAI's ChatGPT and Meta's LlaMA 2. LLMs are trained on extensive text corpora, preprocessed for unsupervised learning, where the model predicts the next word in a sequence using the preceding context. According to Chang et al. (2023), LLMs have significantly increased the potential for achieving artificial general intelligence (AGI), a hypothetical form of intelligence capable of performing any intellectual task comparable to human or animal intelligence, which remains a theoretical concept as of this thesis's writing.

LLMs have demonstrated remarkable versatility across a wide range of domains, from promoting public health, as shown in Biswas (2023), to finance, as shown in Zaremba and Demir (2023). These models have exhibited emergent abilities, where larger models display capabilities not present in their smaller counterparts. The performance enhancement of LLMs significantly correlates with the scaling in terms of computation, model parameters, and training dataset size. This scaling not only betters the performance but also introduces new efficiencies in tasks, including financial sentiment analysis (Chang et al., 2023). LLMs' reasoning capabilities are essential in comprehending complex financial narratives, and their ability to grasp and infer from provided information increases their significance even more. Additionally, LLMs have shown strong performance in sentiment analysis, as evidenced in studies such as Gao, Fisch and Chen (2021), Lopez-Lira and Tang (2023) and Qin et al. (2023).

2.3.2.2 LlaMA 2

LlaMA 2, herein referred to as 'Llama', is a series of open-source LLMs developed by Meta, released to the public on July 18th, 2023. These models, varying from 7 billion to 70 billion parameters, include fine-tuned versions such as LlaMA-2-Chat, which are optimized for dialogue use cases (Touvron et al., 2023). Parameters represent the weights acquired through training, which are then used to predict the subsequent tokens in a sequence (Google, 2023). For context, ChatGPT-3 was trained with 175 billion parameters (Griffith, 2023). The main objective of Llama is "advanced natural language understanding" (Insuasti, Roa and Zapata-Jaramillo, 2023) with applications spanning from conversational AI and content generation to language translation and information extraction.

Llama is built on an optimized auto-regressive transformer architecture, meaning it employs a neural network design (transformer) tailored to generate outputs sequentially (auto-regressive). This architecture has been optimized to enhance performance in language-related tasks. ChatGPT operates on a similar principle, meaning each new word formulated by the model is influenced by the preceding words (Touvron et al., 2023). This feature allows the models to maintain context and coherence in their outputs.

Llama's training involved processing approximately two trillion tokens (units of text) from various public sources, with a focus on data cleanliness and factual accuracy to enrich the model's knowledge base and minimize errors. The fine-tuning process included supervised fine-tuning and reinforcement learning, tailored to align Llama with human preferences, thereby increasing its reliability and safety in practical applications (Touvron et al., 2023). Unlike some closed-source models, such as ChatGPT and BloombergGPT, LlaMA was trained exclusively on public data, predominantly in English, making it particularly suitable for English-language applications (Gal, 2023; Touvron et al., 2023). Additionally, Llama demonstrates exceptional proficiency in NLP tasks, capable of generating highly accurate and human-like texts. The model is distinguished by its ability to understand linguistically complex outputs (Insuasti, Roa and Zapata-Jaramillo, 2023), highlighting its potential in grasping specialized terminology used in financial contexts.

The performance of the different Llama models, especially the fine-tuned versions, has not only surpassed that of many open-source models but also proven to be comparable to some closed-source models, such as ChatGPT, in terms of helpfulness and safety (Touvron et al., 2023). As noted by Insuasti, Roa and Zapata-Jaramillo (2023), Llama's performance suggests that open-source models could potentially replace closed-source models in certain applications. Furthermore, when evaluating Llama against other models, including GPT-3 and GPT-4, Chang et al. (2023) highlighted Llama's competitive edge. Particularly impressive is its ability to tailor sentiment analysis to the specific requirements of industry-related data, a critical feature for conducting detailed financial analyses where subtleties in sentiment can greatly affect interpretations.

Llama's comprehensive capabilities, including contextual understanding, complex language processing, accuracy in sentiment prediction, and efficiency in managing large datasets, make it an invaluable tool in financial sentiment analysis. The model's effectiveness is further corroborated by various studies such as Pavlyshenko (2023) and Breitung, Kruthof and Müller (2023), solidifying its position as a powerful instrument for analyzing financial markets.

In conclusion, Llama, Meta's advanced open-source large language model, emerges as a particularly suitable tool for sentiment analysis in the financial sector. Its optimized auto-regressive transformer architecture, coupled with extensive training on a diverse range of public data, enables Llama to accurately interpret complex financial terminology and contextual nuances, which are essential for analyzing stock market sentiments. The model's proven efficiency in processing large datasets, along with its ability to maintain context and coherence in outputs, makes it invaluable for nuanced sentiment analysis. This capability is highlighted in Breitung, Kruthof and Müller (2023), which demonstrated Llama's ability to extract industry-specific impacts from news headlines.

2.3.2.3 Prompt engineering

Prompt engineering is an emerging area of research denoting the systematic process of formulating, enhancing, and deploying directives or instructions with the aim of directing the output generated by LLMs (Meskó, 2023). With the generality of LLMs, the goal is to tailor LLMs to perform specific tasks as desired (Zhou et al., 2023). Addressing this, recent work has explored fine-tuning (Ouyang et al., 2022; Ziegler et al., 2020), in-context learning (T. Brown et al., 2020) and diverse methods of prompt generation (Gao, 2021) as potential solutions to this challenge and to direct LLMs towards more desirable behaviors. This thesis will follow the approach by Gao, Fisch and Chen (2021).

2.4 Machine learning: Random forest model

The principal objective of machine learning is to identify and codify patterns within unprocessed data. This application can be exemplified by the task of categorizing credit card transactions as fraudulent or legitimate, with the objective of being able to enhance the classification accuracy using historical data (Jordan and Mitchell, 2015). The main advantage compared to, for example Ordinary least squares (OLS), is that ML allows for complex and non-linear relationships (Hastie, Tibshirani and Friedman, 2009).

2.4.1 Decision trees

Decision trees are a machine learning algorithm for both classification and regression (Hastie, Tibshirani and Friedman, 2009). These algorithms excel in managing extensive datasets, enabling the prediction of categorical classes based on training sets and class labels, and are adept at classifying newly acquired data. A decision tree comprises a series of sequential tests, each comparing a numeric feature against a threshold value. Decision trees are primarily utilized for classification purposes in data mining, characterized by nodes representing features in a category and branches signifying the possible values of these nodes. There are three types of nodes, (a) the *root* node, which contains the whole dataset and initiates the split of data into different groups, (b) *internal* nodes, representing possible choices and connecting parent nodes to their children, or *leaf* nodes, and (c) *leaf* nodes, signifying the final outcome of a sequence of decisions, meaning a node without child nodes. Branches symbolize the potential paths stemming from the root and internal nodes, forming a structured model

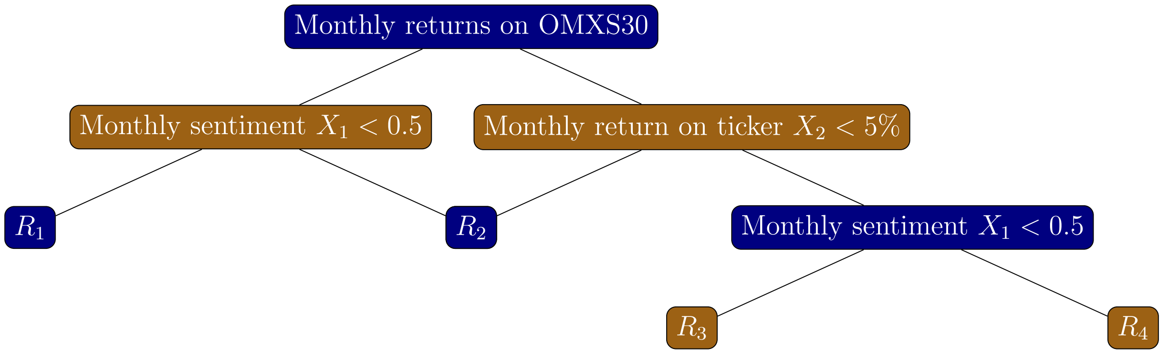

Figure 2.1: Visual representation of the decision-making progress, illustrating the categorization of OMXS30 stocks based on monthly sentiment and closing prices.

where each path from the start to an endpoint signifies a condition, often expressed as "if-then" statements, for making classification decisions.

To illustrate how decision trees function in a practical context, assume there exists a target variable y, which for the purposes of this thesis can be illustrated using the monthly returns on stocks in the OMXS30 index.

$$\hat{f}(X) = \sum_{m=1}^{4} c_m I\{(X_1, X_2) \in R_m\} \tag{2.1}$$

Using two key features, monthly sentiment (X_1) and closing prices (X_2), (2.1) can be thought of as a way to categorize stocks into different performance categories based on the two features. Where $\hat{f}(X)$ depicts the predicted monthly returns, c_m depicts the average return rate for a type of stock in the region, R_m and I depicts an indicator function, meaning it defines a binary condition returning a value of 1 if the condition is true and 0 if the condition is false. The formula divides the stock market into different segments, where each segment represents a combination of sentiment and closing stock return. The tree-based method partitions the stock market into different segments, each representing a unique combination of sentiment and closing price. The tree model then uses these segments to predict the monthly returns, identifying the segment a particular stock falls into based on its sentiment and closing price, and applying the predicted return for that segment. Essentially, the formula aggregates the contributions from each region R_m based on where the input features (X_1, X_2) fall. For each region that the inputs match, the corresponding constant c_m (representing an average return rate for that type of stock) is summed to give the final prediction $\hat{f}(X)$ (Hastie, Tibshirani and Friedman, 2009). This can also be seen in Figure 2.1, the decision tree starts at the root node (Monthly returns on OMXS30), representing the target variable y. From the root, the tree branches out based on certain conditions or thresholds. For example, if the sentiment is below the arbitrary value of 0.5, we follow the left branch starting from the root node, which leads to the leaf node R1. As previously mentioned, the main advantage of choosing decision trees over OLS is if there exist non-linearities, or *heteroscedasticity*, in the dataset. In these cases, OLS will no longer be the Best linear unbiased estimator (BLUE), making decision trees a more viable method as they do not assume a linear relationship between the dependent and independent variables (Hastie, Tibshirani and Friedman, 2009).

2.4.2 Random forest

Bootstrap aggregating, or *bagging*, is a technique that constructs numerous de-correlated decision trees and averages their results, reducing the variance of an estimated prediction function. Random forests combine multiple decision trees, where each tree in the forest relies on the values of a randomly sampled vector, independent and identically distributed across all trees. Random forests, as an ensemble learning method, are particularly suitable for reducing the variance inherent in predictive models, especially those prone to high variability like decision trees (Breiman, 2001). Bagging helps in addressing the overfitting issue often seen in singular decision trees as showcased in Kingsford and Salzberg (2008). Regular decision trees also tend to overfit the training data, especially when there is a substantial amount of noise in the data. As R. Verma and P. Verma (2007) showcases, noise is prevalent in stock market data, making random forests a more viable method rather than singular decision trees. The core idea of bagging is to average multiple noisy but unbiased models, thereby achieving variance reduction without a significant increase in bias. Trees are ideal candidates for bagging since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they benefit greatly from the averaging (Hastie, Tibshirani and Friedman, 2009).

The construction of the model is as follows. For each tree b, ranging from 1 *to* B, a bootstrap sample $Z*$, meaning it randomly selects data from the training set with replacement, of size N is drawn from the training data by sampling with replacement. Each random-forest tree T_b is grown on its bootstrapped dataset $Z*$. It first selects m variables at random from the total p variables, then it identifies the best variables and the optimal split-point among the m selected. It finally divides the node into two daughter nodes using this split point. This process continues until each terminal node of the tree reaches the minimum node size n_{min}. The output of the model is then the ensemble of trees, $\{T_b\}_1^B$ (Breiman, 2001; Hastie, Tibshirani and Friedman, 2009).

$$\hat{f}_{rf}^{B}(X) = \frac{1}{B}\sum_{b=1}^{B} T(x; \theta_b) \tag{2.2}$$

After all B trees have grown, the model takes an average of the predictions from a collection of B individual trees. Each tree contributes its own prediction based on the features and then computes the mean of these predictions according to (2.2), where $\hat{f}_{rf}^{B}(X)$, the target function, represents the aggregate prediction from the entire forest for the input features x. This aggregation process across B trees aims to enhance the predictive reliability and diminish the potential variability that might arise from a single tree's forecast (Hastie, Tibshirani and Friedman, 2009). $T(x; \theta_b)$ is the prediction from an individual decision tree for the input features x, where θ_b are the parameters of the b-th tree. It then adds up the predictions from all individual trees and is then divided by B, the number of trees, to calculate the average prediction across all trees in the forest.

To evaluate the performance of random forests, the dataset is divided into three subsets: training, testing and validation. The training set is used to train the model, meaning the model uses this data to create and train all the decision trees in the forest. Thereafter, it chooses the hyperparameters so that it yields the best model performance based on the validation set. The testing set, which the model has not seen during training, is then used to evaluate the model's performance. This step assesses how well the model generalizes to new, unseen data. The goal is to provide a reliable assessment of the model's performance in real-world scenarios, ensuring it has not just memorized the training data, a concept known as overfitting (Ying, 2019). This is a persistent challenge in machine learning and the underlying issue is that an overfitted model struggles to adapt to elements in the testing set that differ from those in the training set. Three main factors illuminate overfitting, of which the first one is that the model is learning noise in the training set. This happens when the training data is either too small, contains excessive noise or is not sufficiently representative.

To solve the problem of overfitting, some algorithms have been developed to mitigate its impact. One of the solutions is fundamentally the testing in hyperparameter tuning. Appropriately tuned hyperparameters achieve an equilibrium between training accuracy and regularization, thereby mitigating the risks of overfitting. This requires a sufficient amount of samples (Ying, 2019). However, random forests are generally seen to be robust against overfitting because of the employment of bagging (Breiman, 2001; Hastie, Tibshirani and Friedman, 2009). For this thesis, the following key hyperparameters has to be understood:

- **Number of estimators**: This parameter sets the number of trees in the forest.

An increased count can improve the model's accuracy at the expense of higher computational complexity.

- **Maximum depth**: This controls how deep each tree is allowed to grow. Deeper trees can discern more complex patterns, yet there's an increased risk of overfitting.
- **Minimum samples split**: This determines the required minimum number of samples to consider when splitting an internal node. It's a crucial factor for controlling the growth of the trees and mitigating potential overfitting.

To find the optimal hyperparameters that provide the best performance as measured on a validation set, there are two widely used methods; `RandomizedSearchCV` and `GridSearchCV`. `RandomizedSearchCV` selects random combinations of hyperparameters to train the model and outputs the best-performing hyperparameters. `GridSearchCV` is a methodical approach that exhaustively tries every combination of hyperparameters, requiring immense computational power if tested across every existing combination (Scikit, 2023). In the first phase, `RandomizedSearchCV` was used in this paper to navigate through the hyperparameter space. This was followed by the use of `GridSearchCV`, which focused on a refined range informed by the findings from the randomized search. Note that Bayesian optimization, with techniques such as `Hyperopt`, is recognized as state-of-the-art for hyperparameter tuning in many applications (Microsoft, 2023). Nevertheless, for the random forest model built in this thesis involves a smaller number of hyperparameters and limited intervals, the application of `GridSearchCV` and `RandomizedSearchCV` were considered sufficient.

2.4.2.1 Evaluation of results

Some statistical measures are needed to measure the out-of-sample performance. The following two metrics are widely used in random forest modeling (Genuer, Poggi and Tuleau, 2008) and will be employed in the thesis. Furthermore, cross-validation (CV) will be employed to reduce overfitting.

2.4.2.2 Mean square error (MSE)

The Mean square error (MSE) measures the average of the squares of the errors or deviations. It calculates how close a regression line is to a set of points. In this paper, the MSE is used as the loss function during training.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.3}$$

In (2.3), n is the total number of data points, Y_i is the actual (observed) value of the dependent variable for the i-th datapoint and $\hat{Y}_i$ is the predicted value of the dependent variable for the i-th datapoint.

2.4.2.3 R-squared

The R-squared (R^2) metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.

$$R^2 = 1 - \frac{SSR}{SST} \tag{2.4}$$

In (2.4), SSR is the sum of squares of residuals and SST is the total sum of squares.

2.4.2.4 Cross validation

Cross-validation is used to evaluate the generalizability of a model by testing it on multiple subsets of data. The idea is to divide the entire dataset into smaller groups, train the model on some groups, and test it on others. The groups are then scrambled to test and train on different subsets, so the model gets a chance to learn from and be tested on every part of the data. This in turn reduces overfitting.

2.5 Capital asset pricing model

The capital asset pricing model (CAPM) provides estimates of equilibrium expected returns on risky assets by using a linear relationship to connect risk and return. This model assumes rational investors who participate in mean-variance optimisation, have equal investment horizons, and have access to the same information (Bodie, Kane and Marcus, 2021).

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f) \tag{2.5}$$

The formula for CAPM is given by (2.5), where $E(R_i)$ is the expected return of the investment, R_f the risk-free rate, β_i is the volatility of security i, $E(R_m)$ is the expected return of the market and $E(R_m) - R_f$ is the market risk.

Despite a lack of strong empirical data supporting its efficacy, the CAPM is nonetheless commonly used due to its simple framework. This ease of comprehension of the link between predicted risk and return makes it a widespread choice in financial analysis (Johnson, 2015; Reinganum, 1981).

2.5.1 Understanding Jensen's alpha in portfolio performance

After accounting for the beta and average market return of the portfolio, Jensen's alpha measures the average risk-adjusted return of a portfolio that outperforms or underperforms the expected performance as determined by the CAPM (Jensen, 1968). When a portfolio's alpha is positive, it means that it has outperformed the benchmark index and produced returns that are greater than anticipated given its level of risk. A negative alpha, on the other hand, indicates that the portfolio has underperformed in comparison to the benchmark and has delivered lower returns than anticipated given its level of risk. When evaluating the effectiveness of the sentiment-based investment strategy, alpha is one of the critical metrics. Jensen's alpha is also frequently taken into account in performance reviews within the context of a multifactor market model, acknowledging the influence of various market factors on portfolio returns (Bodie, Kane and Marcus, 2021).

$$\alpha = R_p - (R_f + \beta_i(E(R_m) - R_f)) \tag{2.6}$$

Jensen's alpha is calculated with (2.6), derived from (2.5).

2.5.2 Multi-factor models with CAPM

In their papers, Fama and French (1992) and Carhart (1997) recognized key shortcomings in the CAPM. By incorporating multiple factors of risk into the CAPM framework, besides the market risk $E(R_m) - R_f$, they addressed these limitations, paving the way for the development of multi-factor models. This progression resulted in the development of the Fama–French three-factor model (FF3F) and Carhart's four-factor model (C4F). In this thesis, CAPM along with the FF3F and C4F will be utilized to evaluate the dynamic portfolio, constructed with the predicted returns.

$$R_i - R_f = \alpha + \beta_1 \times MKT + \beta_2 \times SMB + \beta_3 \times HMLFF + e \tag{2.7}$$

The FF3F accounts for size, value and market risk and is given by (2.7), where $MKT = E(R_m) - R_f$ is the market risk, as seen in (2.5). SMB (Small Minus Big) is the size risk, the historic excess returns of small-cap companies over large-cap companies. $HMLFF$ (High Minus Low based on Fama–French) is the value risk, historic excess returns of value stocks over growth stocks. The β_i where $i = 1, 2, 3$ represents the volatility of the factors.

$$R_i - R_f = \alpha + \beta_1 \times MKT + \beta_2 \times SMB + \beta_3 \times HMLFF + \beta_4 \times UMD + e \tag{2.8}$$

C4F adds the momentum factor UMD to the FF3F, as seen in 2.8.

The factors, size, value and momentum, represent the systematic risks that the CAPM doesn't take into account. Supporting the size factor, Banz (1981) demonstrated that smaller market capitalization firms have higher risk-adjusted returns than larger firms. Stattman (1980) observed a correlation between a company's book value to market value ratio and expected returns. Furthermore, Jegadeesh and Titman (1993) demonstrated that portfolios favoring stocks with a history of strong performance and shorting those with a history of poor performance could generate excess returns, lending support to the momentum factor.

Empirical evidence, such as in Ericsson and Karlsson (2004) and Johnson (2015), suggests that the incorporation of multiple factors, as proposed by Fama, French, and Carhart, enhances the CAPM's robustness and broadens its scope. The multi-factor models integrate previously overlooked elements, resulting in a more comprehensive knowledge of the many factors impacting stock returns.

3 Methodology

This section explains the methodology used to extract and quantify sentiment from financial news headlines. It discusses the complete process, encompassing data collection, preprocessing, the application of Llama for sentiment analysis, constructing a portfolio based on sentiment, and calculating the alpha. Additionally, details about the model parameters and evaluation criteria are presented. Python and Excel were the primary tools utilized for this analysis. An outline of the methodology can be seen in Figure 3.1 below. For a complete list of Python libraries used, see Table 1 in the Appendix.

An implementation of the methodology can be found at:

https://github.com/vivea123/llama2sentimentanalysis

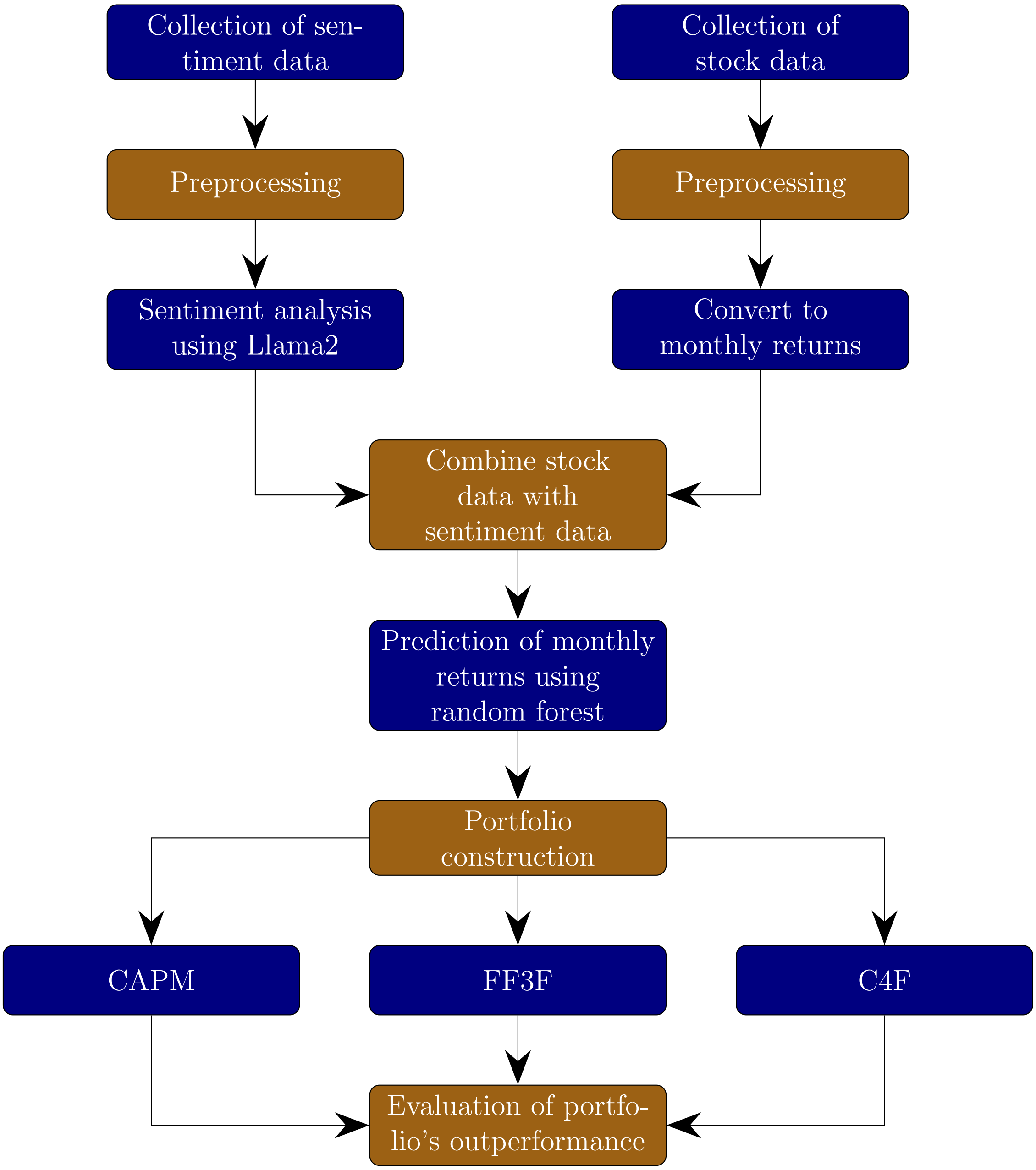

Figure 3.1: Flowchart illustrating an outline of the employed methodology.

3.1 Sentiment data collection and preprocessing

News related to each specific ticker in the OMX Stockholm 30 (OMXS30) was collected from Refinitiv Eikon, herein referred to as Eikon, a database containing financial tools like real-time market data and company-specific news. This database has been widely used in academia (Abinzano, Corredor and Mansilla-Fernández, 2022; Dan Florin and Romascanu, 2021; Esqueda and O'Connor, 2023; Moin, 2023). Company-specific news was defined as the association of individual news headlines to each unique ticker. As an example, Figure 3.1 displays a series of preprocessed news headlines related to four different tickers.

The final dataset curated for this thesis comprised 60,871 unique news headlines, spanning from 2019 to late 2023. The selection criteria for these articles were twofold. Firstly, each article needed to be published in English to maintain consistency in language processing, a significant consideration since Llama was trained predominantly with English texts (Touvron et al., 2023). Secondly, for relevance and specificity, every headline required explicit tagging with the corresponding stock ticker it referenced. This method ensured that the data was pertinent and precisely aligned with the stocks under study.

The time frame selected for this thesis, spanning from 2019 to 2023, was driven by several considerations. Firstly, this period's modernity ensured that the findings remained relevant to the current financial landscape. Secondly, these years were marked by significant macroeconomic events that influenced financial markets, including, but not limited to, the United Kingdom's departure from the European Union, the global COVID-19 pandemic, and the war in Ukraine.

The distribution of the headlines per ticker and per year can be seen in Figure 3.3, and 3.2, respectively. As can be seen, there is quite a large variation in between the number of headlines. The total range for the number of headlines per company spans from 353 for ATCO-A to 7,535 for ABB. Each headline contained unnecessary information not relevant to the analysis such as the exact timestamp the news article was published. The headlines were preprocessed by removing URLs, irrelevant unicode characters, and other non-contributory data, and reformatting them into the desired format.

Table 3.1: Examples of news headlines, its corresponding tickers, and date published.

Ticker	Date	News headline
ABB	2023-11-22	ABB gets $547 Mln Loan From the European Investment Bank
ASSA-B	2022-12-02	Assa Abloy locks in $800 mln deal to sell Emtek, Yale businesses to Fortune Brands
ELUX-B	2023-08-02	ELECTROLUX ELUXb.ST: JP MORGAN RAISES TARGET PRICE TO SEK 113 FROM SEK 100
HM-B	2022-12-05	H & M Hennes & Mauritz AB (publ) (OTCMKTS:HNNMY) Stock Rating Reaffirmed by Morgan Stanley

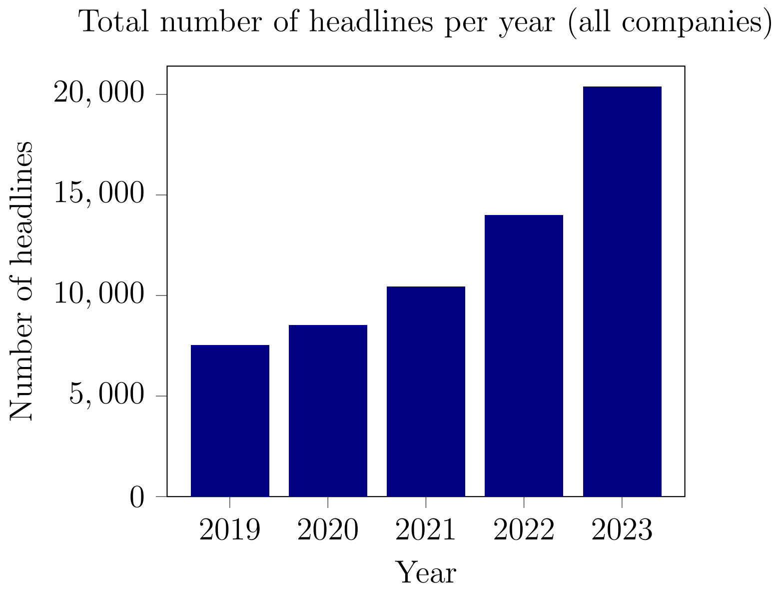

Figure 3.2: Total number of headlines per year for all companies.

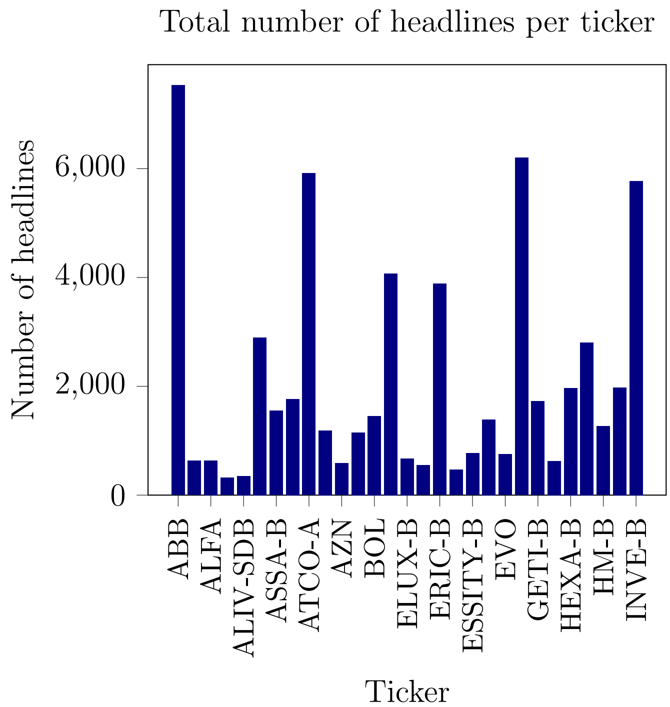

Figure 3.3: Total number of headlines per ticker.

A primary limitation of this thesis was the exclusive use of English texts in analyzing the OMXS Stockholm 30 (OMXS30). The Swedish financial market is influenced by both global and local economic events and sentiments. While English is a dominant language in global finance, local investor sentiment, which can significantly impact the index, is likely captured in native Swedish media and texts. By limiting the analysis to English texts, this thesis might have overlooked key sentiment indicators present in Swedish sources, potentially leading to a skewed or incomplete understanding of the market sentiment affecting the OMXS30.

The tickers included in our analysis were static, representing the index composition as of December 2023. This focus had limitations, as it might not account for the historical sentiment and performance of companies no longer part of the index, see Table 2 in Appendix for leavers and joiners in the index. The chosen collection of tickers might not have fully encapsulated historical market dynamics, potentially skewing the analysis toward better-performing companies. However, this approach aligned with the current state of the Swedish stock market, offering utility for current market analysis.

Another limitation was the sole reliance on news headlines for sentiment analysis. Headlines, while concise and impactful, may not always provide the complete context or depth required to accurately gauge market sentiment (Tetlock, 2007). They are designed to capture attention and might be sensationalized, thus not always reflecting the true sentiment of the accompanying article or the broader market (Soroka, Fournier and Nir, 2023). This reliance could lead to an overemphasis on short-term reactions or extreme sentiments, potentially missing nuanced or long-term trends. However, similar studies have exclusively used news headlines for sentiment analysis with satisfactory results (Breitung, Kruthof and Müller, 2023; Colacicchi, 2022; Fatouros et al., 2023; Lopez-Lira and Tang, 2023).

3.2 Stock price data

OMXS30, which encapsulates the 30 most actively traded stocks on Nasdaq Stockholm (formerly Stockholm Stock Exchange), presented a unique case for sentiment analysis within the field of finance. As previously mentioned in section 2.2, and as illustrated by Baker and Wurgler (2007), investor sentiment was acknowledged as a pivotal factor influencing stock returns. This underlines the investor sentiment as particularly relevant in understanding the dynamics of the Swedish stock market.

Given the OMXS30's position as a less explored index compared to global giants like the S&P 500, it offers a distinctive platform to study market dynamics and the effectiveness of sentiment analysis in smaller markets. This thesis not only contributes to the existing body of knowledge on the Swedish stock market but also explores the broader applicability and potential of sentiment analysis in finance.

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{3.1}$$

The stock prices for each of the 30 tickers were retrieved from Yahoo Finance, ranging from 2019 until late 2023. To see the companies that were included in the full dataset, see Table 3 in Appendix. In order for the news to be fully incorporated into the stock prices, this thesis used monthly prices as investors are empirically shown to be slow reactors to news (Bodie, Kane and Marcus, 2021). We used the closing price for each of the 30 tickers, so preprocessing involved removing the trading volume and opening price. Furthermore, the monthly returns were calculated using (3.1), where P_t is the closing price at time t, P_{t-1} is the price at time $t-1$ and r_t is the return at time t.

3.3 Sentiment analysis methodology

For the annotated sentiment on news headlines, Llama was employed. The chosen model was `llama-2-7b-chat.Q4_K_M`, the smallest available model. Touvron et al.

(2023) found that the performance of all Llama models were significantly superior to randomness and outperformed all tested open-source models while being comparable to closed-source models. For this thesis the chosen model, albeit the smallest, was therefore deemed sufficient. A larger model would require a larger computational power. Similar to Breitung, Kruthof and Müller (2023) this thesis adopted three sentiment categories for annotation: positive, negative, and neutral.

The key variables used in the Llama model setup can be seen in Table 4 in Appendix. `Temperature` is a parameter that controls the randomness of Llamas output. A higher temperature will result in more creative responses, while a lower temperature will result in more accurate and factual text. The temperature was therefore set to 0 to maximize the factual accuracy. The `top_p` parameter balances the uniqueness of the generated text while maintaining readability and logical flow (Horton et al., 2021). A higher `top_p` increases the likelihood of the model choosing less probable outputs. The motivation behind setting `top_p` to 0.5 was to strike a balance in the probability of the produced output between highly likely and less likely tokens.

As explained in section 2.3.2.3, prompting large language models (LLM) is critical for obtaining specific outputs. This thesis used a transformed version of the best overall performing prompt, named "GPT-P4", from the study by Fatouros et al. (2023). Since their study focused on the Forex market, the prompt was adapted to fit the scope of this thesis. Furthermore, the prompt had to be transformed in a way that corresponded to Llamas's syntax of prompting, for details refer to code block 1 in Appendix. Additionally, zero-shot prompting approach was employed, meaning no prior fine-tuning of the model was conducted. The reasoning for this stems from the desire to assess the model's capabilities in its most general form. This approach allows for an evaluation of the model's out-of-the-box applicability as well as taking on a similar approach to that of Fatouros et al. (2023), as they also used a zero-shot prompting approach.

$$Weighted\ sentiment = \frac{\sum_{i=1}^{n} s_i \lambda^{(n-i)}}{\sum_{i=1}^{n} \lambda^{(n-i)}} \tag{3.2}$$

The model processed each headline sequentially, categorizing sentiments into numerical values: positive as 1, neutral as 0, and negative as -1. Since stock prices were provided monthly and the headlines were published sporadically, the sentiment annotations from the headlines were weighted using (3.2) and then aggregated on a monthly basis. In this context, s_i denotes the numeric sentiment values, λ is the decay factor, and n is the number of sentiment values. The formula uses an exponential decay factor, a value between 0 and 1, that reduces the weight of past sentiments, with a higher factor indicating a slower decay reduction and thus more weight to earlier observations. Considering the relevance of historical news in impacting present stock prices, as highlighted in Boudoukh et al. (2012), and the absence of a definitive method for assessing the precise impact of previous news items on present stock returns, this thesis adopted a decay factor of 0.95. This choice is aligned with Diebold, Hahn and Tay (1999) in their analysis of multivariate density forecasting. The exponentiality ensures that the most recent data points have more influence on the outcome than older ones. For example, if the decay factor is equal to 0.9, the most recent sentiment score would have a weight of 0.9^{n-1}, the next one would have 0.9^{n-2}, and so on. In essence, the older the data point, the smaller the weight.

3.4 Prediction model: Random forest

This section of the thesis delineates the methodology employed to predict stock returns using a random forest (RF) model, intertwining with the theoretical framework provided by Hastie, Tibshirani and Friedman (2009). The model employed a matrix consisting of 58 features, 29 sentiment features and 29 stock return features, in order to predict the target variable, and split the data into validation, testing, and training sets. The focus on random forests stemmed from their proven efficacy in handling complex, high-variance datasets, a characteristic prevalent in financial markets, as discussed in section 2.4.2.

The target variable was the daily returns for each specific ticker, while the features comprised the monthly sentiment and monthly returns for each stock, resulting in a matrix with 58 columns. The data was partitioned, or *split*, into validation, testing, and training sets, allocating 60% for training, 10% for validation, and 30% for testing, in accordance with the methodology advocated by Xu and Goodacre (2018). Time series data, such as stock returns and headlines, exhibit temporal dependence, meaning their current values are influenced by past values. Conventional random data splitting methods can disrupt this chronological order, leading to inaccurate predictive models, as demonstrated by Tashman (2000). To preserve the temporal sequence, the `TimeSeriesSplit` method was employed for partitioning our train-test split, as illustrated in Figure 3.4. The vertical axis represented each of the cross-validation folds, while the horizontal axis displayed the entire dataset. Additionally, a validation set was introduced within each training phase for fine-tuning hyperparameters, taken from the end of each training set. Consequently, in each split, the model was trained on the initial portion of the data, validated on the subsequent portion, and finally tested on the following section. With each subsequent iteration, or *split*, the

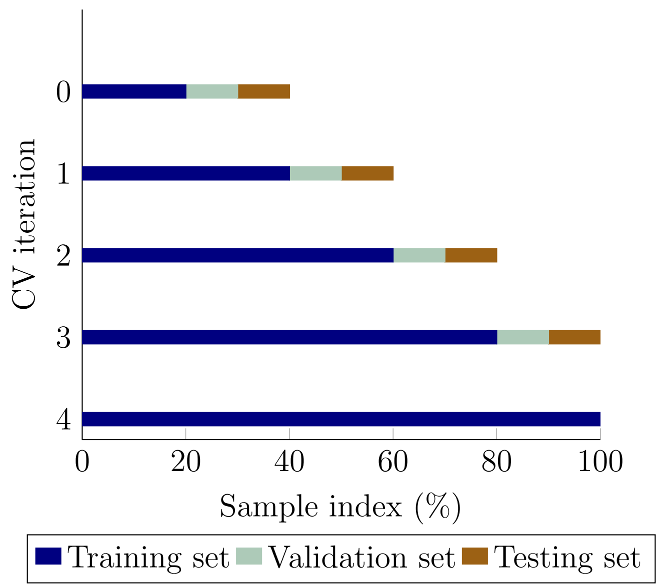

Figure 3.4: Visualization of time series split with training, validation, and testing sets, each bar, or split. For each split, the training set grows to include all available data up until the start of the validation set and thereafter tested on the subsequent portion of data.

training set expanded to include all available data up until the start of the validation set. This approach ensured the model was trained on historical data and then tested on subsequent, unseen data, as depicted by each bar, or *fold*, in the figure. The training dataset progressively included more observations, while the test set advanced, each representing 20% of the total data. Finally, preventing the model from accessing future data during training was paramount to avoid sequential data disruption and look-ahead bias, which is why this method was employed.

3.4.1 Feature engineering

A matrix encapsulating both closing price and weighted sentiment scores for each ticker was constructed, resulting in a matrix of 58 columns. Atlas Copco (ATCO) has both their A and B share in the OMXS30 index, hence the matrix comprised 58, not 60, features. The sentiment scores were derived by employing the LLM from Meta, Llama, and then converted from qualitative sentiment data into quantifiable scores. This matrix design is predicated on the hypothesis, supported by Brown and Cliff (2004), that there is a significant correlation between investor sentiment and stock returns. Each row in the feature matrix was shifted back one period (one month), also known as *lagging* the features, to train and predict the model stock returns for $t + 1$. The model fitting process involved minimizing the MSE during the training process, meaning the algorithm adjusted the model parameters to reduce the MSE, i.e. the differences between the predicted and the actual values. R^2 was utilized to measure

Table 3.2: Hyperparameter configuration for the random forest regressor.

Parameter	Value
`n_estimators`	100
`bootstrap`	True
`max_depth`	6
`max_features`	"sqrt"
`min_samples_leaf`	4
`min_samples_split`	10
`random_state`	42

the goodness of fit.

Hyperparameter tuning was performed through a sequential two-step process. Initially, `RandomizedSearchCV` was conducted to efficiently explore a broad range of hyperparameters. This was followed by a more focused `GridSearchCV`, which meticulously fine-tuned the parameters within the narrowed scope identified by the randomized search. During both steps, `TimeSeriesSplit` was utilized within the `RandomizedSearchCV` and `GridSearchCV` to ensure that the validation process respected the temporal order of the data. The final hyperparameters can be seen in Table 3.2. To evaluate the predictions, the model was run again, this time excluding all the sentiment features but using the same hyperparameters.

3.5 Portfolio construction based on SA

This section explains the construction of a dynamic investing portfolio based on sentiment-based predicted returns of the stocks in the OMXS30 index. The main objective was to determine if the portfolio displayed a significant alpha, indicative of market outperformance, when evaluated against the CAPM, FF3F, and C4F.

3.5.1 Dynamic portfolio construction

The portfolio was constructed by systematically selecting stocks from the OMXS30 index based on their sentiment-driven predicted returns at the start of each period t, with t representing a month. Long positions were assigned to the six stocks with the highest predicted returns, while short positions were allocated to the six stocks with the lowest predicted returns. The long-short strategy, substantiated by the insights from Jacobs, Levy and Starer (1998), sought to leverage market fluctuations for both rising and falling stocks, diversifying the portfolio, providing a balanced approach to risk and return. The strategy was implemented for the out-of-sample period from

March 31, 2020, to November 30, 2023, as the in-sample period leading up to March 31, 2020 was dedicated to training the random forest model.

The performance of the portfolio was evaluated during the subsequent period, denoted as $t + 1$. The evaluation took into account both the gains from long positions and outcomes from short ones. The portfolio was reassessed at the end of each period $t + 1$, re-ranking the companies based on the most recent projections. This procedure ensured that the portfolio was consistently updated with the top and bottom six stocks for each new period, maintaining the strategy of equal-weighted long and short positions throughout the investing term.

$$r_{p,t} = \sum_{k=1}^{n} w_{k,t} r_{k,t} \tag{3.3}$$

To calculate the returns of the portfolio after each rebalancing period, (3.3) was used, where $r_{p,t}$ represents the portfolio's total return for month t, and $w_{i,t}$ is the weight of the i-th stock in the portfolio for month t. The variable $r_{i,t}$ is the return of the i-th stock for month t. n is the total number of stocks in the portfolio.

3.5.2 Portfolio evaluation: CAPM, FF3F and C4F

The sentiment-based portfolio's performance was assessed against the benchmarks set by CAPM, FF3F, and C4F. To establish the portfolio's relationship to the market, the portfolio's returns were regressed against market returns, specifically the OMSX30 index returns sourced from Yahoo Finance. The monthly U.S. Treasury bill risk-free rates from the AQR (2023) dataset were used in (2.5), (2.7) and (2.8).

The goal was to assess whether the portfolio's alpha, or its outperformance over the market, remained constant after accounting for various factors affecting stock returns. This analysis aided in determining whether the observed positive alpha was attributable to a distinct sentiment risk factor, or if it could be explained by well-established market factors. A monthly average of Jensen's alpha was calculated over the entire time period with (2.6) for all models. This was done to compare the performance to the OMXS30 index while taking the respective risks (market, size, value, and momentum) into account. Factors from the monthly AQR (2023) dataset, specifically for the Swedish market, were incorporated. These comprise MKT (market risk), UMD (up minus down), HML FF (high minus low based on Fama–French), and SMB (small minus big).

If alpha was significant in these models, it would indicate the presence of a distinct sentiment risk factor. If, on the other hand, alpha became insignificant, the analysis would determine which specific factor(s) were responsible for it. This would disclose if the performance of the portfolio was attributable to exposure to established risk factors or a unique sentiment risk factor.

4 Results

This thesis examined the integration of sentiment analysis into a random forest model for predicting stock returns. Utilizing Meta's LlaMA 2 for sentiment analysis, news headlines were categorized as positive, negative or neutral. The sentiments were further numerically quantified and aggregated on a monthly basis. The predictive model encompassed 58 features; 29 were the monthly returns for each stock, and the other 29 were the weighted sentiment scores. A dynamic long-short portfolio was built with the predicted returns and then evaluated against CAPM, FF3F, and C4F to assess the performance of the portfolio over the OMXS30 index market benchmark.

4.1 Predictions

To encapsulate the role sentiment plays, some measures were used to compare the predictions between the random forest model with and without the sentiment features. These findings can be found in Table 4.1, which shows an average of the MSE and R^2 for all $n = 5$ splits in `TimeSeriesSplit`, as well as the correlation when the model was run with and without the sentiment features. The same hyperparameters were used when running the random forest with and without sentiment. The table showcases this, as the fit of R^2 and correlation shows a much better result when including sentiment features. This result echoes the findings of previous literature like those of Bollen, Mao and Zeng (2011) and Shapiro, Sudhof and Wilson (2020), which established a positive correlation between investor sentiment and stock returns. Additionally, it is also in line with Fatouros et al. (2023) which showcased that by exploiting LLMs, one can predict stock returns for subsequent periods. Finally, the findings align with the behavioral finance perspective, as highlighted by Shiller (2000), suggesting that media-influenced investor sentiment can shape stock price expectations. Moreover, our use of the random forest model for prediction resonates with the sentiment analysis methodology outlined in Ferreira, Gandomi and Cardoso (2021), and demonstrates the utility of modern AI and machine learning techniques in financial prediction, a concept further supported by the study of Huang, Capretz and Ho (2021). The findings also challenge the traditional views of the EMH, as discussed in Fama (1970) and Malkiel (2003), by providing empirical evidence that supports the feasibility of stock prediction using sentiment analysis, a factor traditionally overlooked in major asset

Table 4.1: Measures comparing the random forest model, run with and without sentiment features.

	With sentiment			Without sentiment		
Ticker	**MSE**	R^2	**Corr.**	**MSE**	R^2	**Corr.**
ASSA-B	0.0007	0.5143	0.8524	0.0022	-0.0970	-0.1821
ESSITY-B	0.0014	0.5838	0.6226	0.0040	0.0269	0.0618
TELIA	0.0013	0.6508	0.6704	0.0016	-0.0943	-0.1711
INVE-B	0.0003	0.7588	0.7828	0.0010	0.0856	0.1093
SINCH	0.0367	0.4822	0.6390	0.1138	-0.0903	-0.2475
GETI-B	0.0071	0.3132	0.6097	0.0024	-0.1020	-0.0753
SWED-A	0.0024	0.7070	0.7416	0.0031	-0.0124	0.1412
SKF-B	0.0014	0.7387	0.7456	0.0052	0.0232	-0.0251
TEL2-B	0.0020	0.6226	0.7161	0.0054	-0.1609	-0.0669
SCA-B	0.0010	0.2228	0.5283	0.0006	-4.9472	-0.2324
NDA-SE	0.0021	0.5543	0.6300	0.0006	-2.2029	-0.1368
ATCO-A	0.0025	0.3212	0.7775	0.0029	0.1695	-0.0456
AZN	0.0009	0.3415	0.5311	0.0031	-5.6630	-0.1045
VOLV-B	0.0006	0.8032	0.7042	0.0025	0.0494	-0.0560
SEB-A	0.0008	0.7462	0.5771	0.0019	-0.2614	-0.0235
SHB-A	0.0025	0.5537	0.5944	0.0017	-0.4164	0.0044
ALIV-SDB	0.0027	0.2609	0.6169	0.0059	-0.5685	0.0479
KINV-B	0.0008	0.8398	0.6277	0.0096	-1.1363	-0.0964
ABB	0.0014	0.5437	0.7571	0.0037	0.0411	-0.0462
ELUX-B	0.0066	0.5039	0.7062	0.0105	-0.2351	-0.1519
BOL	0.0032	-0.1164	0.5206	0.0040	-0.1976	-0.0092
HM-B	0.0117	0.3971	0.6155	0.0042	0.1892	-0.1640
NIBE-B	0.0063	-2.3041	0.6562	0.0167	-10.0135	0.2737
EVO	0.0028	0.4522	0.6656	0.0124	-1.4339	-0.1480
ALFA	0.0005	0.7168	0.6389	0.0028	-0.4035	-0.1129
SAND	0.0005	0.8633	0.8345	0.0036	-0.2778	-0.0116
SBB-B	0.0568	0.2127	0.6880	0.0328	-0.1763	-0.0326
HEXA-B	0.0059	0.4269	0.6876	0.0155	-0.0274	-0.0192
ERIC-B	0.0013	0.5955	0.3335	0.0043	-0.1292	-0.2525

pricing models like CAPM, FF3F, and C4F.

Figure 4.1 shows the actual versus predicted returns over time, with and without sentiment features in the random forest model for Assa Abloy (ASSA-B). This shows that the model consistently underestimates the actual returns but seems to, most of the time, follow the same direction as the actual returns. The training was conducted during the initial period, commencing in June 2020.

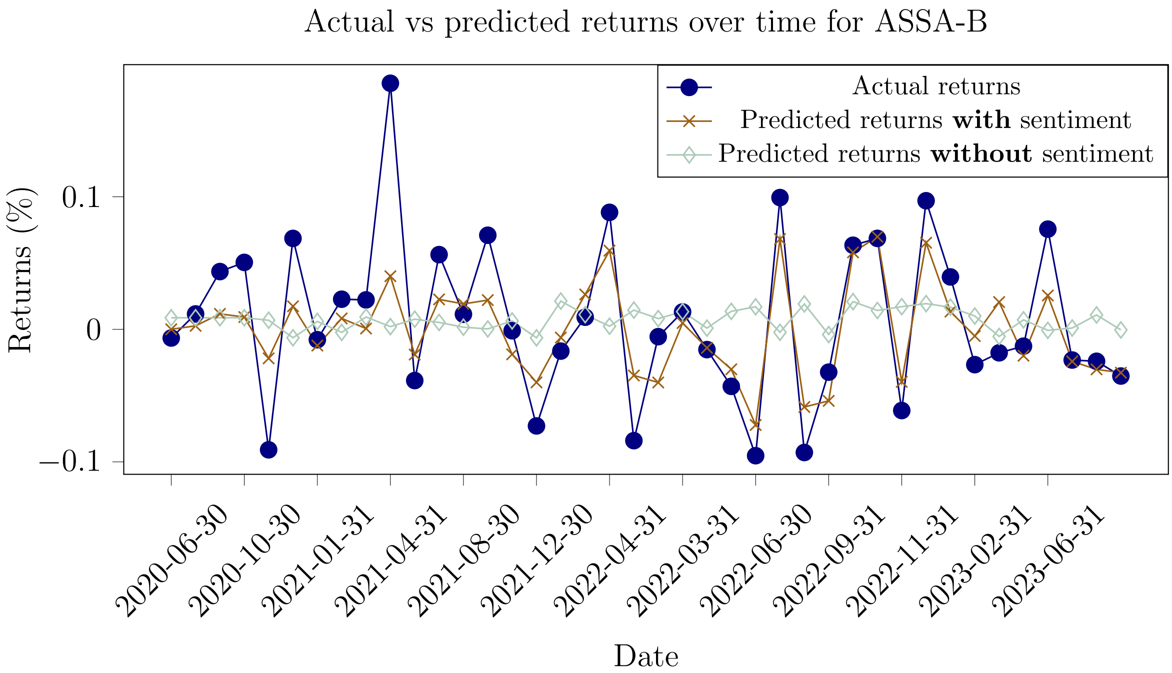

Figure 4.1: Visualization of actual and predicted returns over time, with and without sentiment for ASSA-B.

4.2 Constructed portfolio

Using the predicted returns outlined in section 4.1, the dynamic portfolio was evaluated against the CAPM, FF3F, and C4F to evaluate whether the constructed portfolio could outperform the OMXS30 index and to assess if sentiment is a risk factor.

As seen in Table 4.2, CAPM achieved a significant(***)[1] monthly average alpha of $\alpha =$ 8%, with a p-value of 0.002, suggesting systematic risk cannot explain the portfolio's abnormal returns. These findings align with the research conducted by Boido and Fasano (2014), who also investigated the interplay between sentiment and stock returns within a CAPM framework and also demonstrated that the inclusion of sentiment factors, when evaluated against CAPM, could produce excess returns. Furthermore, the low R^2 in the CAPM, suggests that the market factor alone does not fully explain the portfolio's return variance. This is consistent with findings by Fama and French (1992) and Carhart (1997), who introduced FF3F and C4F to better account for returns.

In the FF3F model, alpha remains remains significant (**)[1] at a monthly average of $\alpha = 8.07\%$ with a p-value of 0.03, even after introducing size and value factors. This suggests that the outperformance of the portfolio is not entirely explained by these additional factors, hinting that other factors, potentially including sentiment, are in

[1]*, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively

Table 4.2: OLS regression results comparison for the CAPM, FF3, and C4F models.

Metric	CAPM	FF3F	C4F
Observations	45	45	45
R-squared	0.0053	0.0349	0.1313
Alpha	0.0800	0.0807	0.0797
P-value of Alpha	0.002	0.03	0.01
Beta Coefficients			
Market Premium (MKT)	0.0208	0.0188	0.0014
Size (SMB)	-	0.0778	0.0907
Value (HML)	-	-0.0110	0.0121
Momentum (UMD)	-	-	0.1073
Durbin-Watson	1.6080	1.7150	1.5617

effect.

Lastly, in the C4F, alpha is still significant (**)[1] at a monthly average of $\alpha = 7.97\%$ with a p-value at 0.01, suggesting that the portfolio continues to outperform even after factoring for market, size, value, and momentum characteristics. The significance (**)[1] of the momentum factor with a UMD coefficient of 0.1073 shows that momentum risk accounts for a portion of the portfolio's abnormal returns.

In the study by Lopez-Lira and Tang (2023), a long-short portfolio strategy was employed, using an LLM, mirroring the methodology used in this thesis. Their findings revealed that this strategy led to a portfolio performance that was markedly superior compared to the market portfolio. To gauge the extent of their portfolio's outperformance, Lopez-Lira and Tang (2023) applied the Sharpe ratio, discovering that their strategy yielded a significantly higher Sharpe ratio relative to other strategies. Although the use of the Sharpe ratio was not explored in this thesis, the method of assessment and the resultant conclusions align closely with the findings of this thesis, providing further support that exploiting LLMs for sentiment analysis can yield excess returns. Conversely, Lopez-Lira and Tang (2023) did not benchmark their portfolio to traditional asset pricing models, a factor that elevates a distinction of this thesis.

Finally, in the study by Xing, Cambria and Welsch (2018a), the integration of sentiment analysis led to improved portfolio performance. Xing, Cambria and Welsch (2018a) found that their approach improved the annualized return by approximately 10% on average, and also enhanced the Sharpe ratio. These results further underscore the importance of qualitative data in financial prediction, resonating with the views of Shiller (2000, 2003) on the limitations of efficient market theories and the role of behavioral factors.

5 Discussion

This thesis's exploration into the integration of sentiment analysis with stock return predictions, particularly for the OMXS30 index, reveals that the inclusion of sentiment-based features can explain stock returns. The adoption of Meta's LlaMA 2 model for sentiment analysis has notably augmented the predictive power of the random forest model. This enhancement is not just a theoretical improvement but is empirically evidenced by the superior performance of the sentiment-informed portfolio over the OMXS30 index market benchmark. Such a result is pivotal, as it demonstrates the benefits of incorporating qualitative insights into quantitative financial analyses. Additionally, this thesis aimed to explore the potential of investor sentiment as a missing risk factor in financial market analysis, particularly in the context of the EMH, CAPM, FF3F, and C4F. The outcome suggests that with the addition of sentiment as a risk factor, models like the CAPM, FF3F, and C4F could potentially enhance their ability to explain portfolio returns.

Table 4.1 demonstrates that incorporating sentiment features noticeably enhances the prediction accuracy, reflected by a reduction in MSE, an improved R^2 fit, and a rise in correlation. While BOL and NIBE-B are exceptions, exhibiting a negative R^2, this anomaly could be due to factors specific to each company. Despite NIBE-B and BOL having 464 and 1,550 unique headlines respectively, these figures suggest that the quantity of headlines does not guarantee improved results. The potential impact of the qualitative nature of these companies' headlines on their results, however, was not further investigated. Moreover, the model without sentiment features mainly exhibits negative R^2 values, suggesting the difficulty of predicting the subsequent month's return based solely on the current month's data. Such a task would dis-align with the weak form of efficiency as imposed by the EMH. Enhancing the models with other factors, such as a simple moving average, or a relative strength index indicator in the manner of what Park and Irwin (2007) discussed, could potentially offer better predictions without sentiment features. All the same, the table illustrates the positive impact of integrating sentiment-based elements into the predictions, with noticeable improvements across the board. Lastly, the variability in the predictions for different tickers by the random forest model suggests that sentiment's influence varies depending on the company.

The low R^2 in CAPM, as shown in Table 4.2, indicates that the market factor alone does not explain much of the portfolio's return variance, however, with the inclusion of additional factors in FF3 and C4F increase the explanatory power of these models. And yet, the fact that the portfolio still outperforms the market when evaluated against these asset pricing models, give rise to the potential that the sentiment-based strategy is capturing unique aspects of stock returns. Approaching cautiously, it is important to consider the potential for a spurious relationship among some variables, hinting at the existence of another, unseen factor that might be influencing the results. Additionally, the Durbin-Watson values are close to 2 for all models, indicating no significant autocorrelation. This implies that the returns are not heavily influenced by their past values, and a momentum trading strategy might not be viable for this specific portfolio. Overall, the sentiment-based portfolio adds value to the portfolio, as evidenced by the significant alpha across all models. The strategy is successful in generating excess returns above what traditional risk would predict.

We believe that the implications of this thesis go beyond its immediate findings. It emphasizes the rising importance of qualitative data, such as investor sentiment, in financial prediction. This trend towards a more integrated approach calls into question the previous emphasis on solely quantitative measurements and opens up new opportunities for thorough market analysis. Furthermore, the effective implementation of a large language model such as LlaMA 2 in this specific financial context demonstrates the adaptability and promise of such technologies in specialized domains.

5.1 Limitations

This thesis, while providing valuable insights into the relationship between investor sentiment and stock market performance, has several limitations. One notable limitation is its emphasis on the OMXS30 index. While this allows for a thorough examination of the Swedish market, it may fall short of capturing the numerous and complicated behaviors observed in larger, more diversified foreign markets. Furthermore, the emphasis on English language sources in sentiment research adds a possible bias, perhaps disregarding non-English views that might be crucial in the OMXS30 market.

Furthermore, the thesis's reliance on a single advanced language model, may not adequately represent the different possibilities of contemporary NLP technology. This constraint casts doubt on the research's generalizability across diverse linguistic models and financial settings. Additionally, while the random forest model has shown to be beneficial, it may not capture the complexities and insights that other sophisticated prediction models may provide. The evaluation measures used in this work, notably MSE and R-squared, represent a standard approach to model evaluation. However, these measurements may not fully cover all aspects of prediction accuracy or model robustness. Additionally, this thesis assumed a nonexistence of transaction costs, a limitation that impacts the findings of this thesis as there are real costs in trading. Finally, there is a marginal probability that LlaMA 2 was trained on the data upon which this thesis based its analysis on. On one hand, seeing as Refinitiv Eikon is close sourced, this might not be a problem but on the other hand, the information contained within the news headlines may be present in public sources. Together, these limitations highlight areas for future research, notably in diversifying the analytical tools and methods for a more holistic approach to financial market prediction.

5.2 Future research

The potential for expanding this research is wide and multifaceted. To begin, widening the scope to include stocks other than those in the OMXS30 index could offer valuable insights. Exploring stocks from various sectors or market capitalizations may show a variety of market behaviors and investor sentiment patterns. Future research could delve into why the influence of sentiment varies across different companies, and if it varies across sectors as well. Other research could also expand upon the findings of this research, with the inclusion of transaction costs in order to evaluate a more real-world application.

Additionally, incorporating alternative prediction models could further enhance the analysis's robustness. While the random forest model gave significant insights in this thesis, testing with other well-known machine learning approaches like support vector machines (SVMs), artificial neural networks (ANNs), or gradient boosting machines (GBMs) might reveal additional levels of prediction accuracy. A study by Xing, Cambria and Welsch (2018a) explored asset allocation using deep-learning models, including convolutional neural networks, demonstrated this potential. Moreover, research comparing various machine learning models, such as random forests, LSTMs, ANNs, SVMs for stock market trend prediction further supports this approach, as seen in Huang, Capretz and Ho (2021) and Soong and Tan (2021).

Moving beyond MSE and R-squared to include additional metrics such as mean absolute error (MAE), root mean square error (RMSE), or advanced Bayesian approaches would give a more comprehensive assessment of model performance. The utility and contexts of using MAE, RMSE, and Bayesian approaches are discussed in depth in studies like Gandhmal and Kumar (2019) and Ericsson and Karlsson (2004). Finally, assessing portfolio outperformance using other financial metrics besides Jensen's alpha, such as the Sharpe ratio, as Xing, Cambria and Welsch (2018a) utilized in their study, Sortino ratio, or benchmark comparisons may provide a more comprehensive understanding of the portfolio's relative success in the context of market risks and volatility. These additions would not only conform with conventional financial procedures but would also provide a larger and more in-depth review of sentiment analysis in stock market prediction.

6 Conclusion

This thesis aimed to resolve two primary questions: First, whether investor sentiment is a reliable predictor of stock returns for OMXS30 index companies. Second, whether a portfolio built on these predicted returns can produce excess return. We also discuss the potential of sentiment being a missing risk factor in traditional asset pricing models. Our research reveals that stock returns can be accurately predicted by exploiting investor sentiment. Furthermore, the portfolio based on these predictions notably exceeds market performance, as evidenced by a positive alpha, especially when compared to conventional models like CAPM, FF3F, and C4F. Moreover, this thesis demonstrated that sentiment is a missing factor in the conventional asset pricing models.

This thesis took a different approach to that of other papers, as we investigated the Swedish stock market as well as using the open-source LLM, LlaMA 2. The effective incorporation of qualitative data derived from LlaMA 2 into the random forest algorithm explored the potential of financial sentiment analysis, offering a nuanced perspective of OMXS30 market dynamics by highlighting the influence of sentiment in stock market performance.

Extending the scope of sentiment analysis to include other languages than English would offer a more comprehensive picture of global investor sentiment. Further research comparing various language models in stock return prediction might provide insights into the best tools for such assessments. Furthermore, employing more complex machine learning techniques might help to refine and improve the prediction of stock market models.

In short, we hope that our contributions will spur further research in the connection between LLMs and financial markets. The application of LLMs within the financial markets remains in a preliminary stage, and we await with interest the advancements that future publications will present. Overall, our findings demonstrate that it is possible to generate excess returns when evaluating against conventional asset pricing models by leveraging LlaMA 2 in the expanding field of sentiment analysis.

Bibliography

Abinzano, Isabel, Pilar Corredor and José Manuel Mansilla-Fernández (1st June 2022). "Sovereign Debt Holdings and Banks' Credit Risk: Evidence from the Eurozone". In: *Finance Research Letters* 47, p. 102600. ISSN: 1544-6123. DOI: 10.1016/j.frl.2021.102600. URL: https://www.sciencedirect.com/science/article/pii/S154461232100502X (visited on 21/12/2023).

AQR (2023). *Betting Against Beta: Equity Factors Data, Monthly.* URL: https://www.aqr.com/Insights/Datasets/Betting-Against-Beta-Equity-Factors-Monthly (visited on 06/01/2024).

Araci, Dogu (27th Aug. 2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.* arXiv: 1908.10063 [cs]. URL: http://arxiv.org/abs/1908.10063 (visited on 21/12/2023). preprint.

Baker, Malcolm and Jeffrey Wurgler (1st Apr. 2007). "Investor Sentiment in the Stock Market". In: *Journal of Economic Perspectives* 21.2, pp. 129–151. ISSN: 0895-3309. DOI: 10.1257/jep.21.2.129. URL: https://pubs.aeaweb.org/doi/10.1257/jep.21.2.129 (visited on 22/12/2023).

Banz, Rolf W. (1st Mar. 1981). "The Relationship between Return and Market Value of Common Stocks". In: *Journal of Financial Economics* 9.1, pp. 3–18. ISSN: 0304-405X. DOI: 10.1016/0304-405X(81)90018-0. URL: https://www.sciencedirect.com/science/article/pii/0304405X81900180 (visited on 03/01/2024).

Barberis, Nicholas and Richard Thaler (1st Jan. 2003). "Chapter 18 A Survey of Behavioral Finance". In: *Handbook of the Economics of Finance*. Vol. 1. Financial Markets and Asset Pricing. Elsevier, pp. 1053–1128. DOI: 10.1016/S1574-0102(03)01027-6. URL: https://www.sciencedirect.com/science/article/pii/S1574010203010276 (visited on 03/01/2024).

Bird, Graham, Wenti Du and Thomas Willett (1st Apr. 2017). "Behavioral Finance and Efficient Markets: What Does the Euro Crisis Tell Us?" In: *Open Economies Review* 28.2, pp. 273–295. ISSN: 1573-708X. DOI: 10.1007/s11079-017-9436-1. URL: https://doi.org/10.1007/s11079-017-9436-1 (visited on 03/01/2024).

Biswas, Som S. (1st May 2023). "Role of Chat GPT in Public Health". In: *Annals of Biomedical Engineering* 51.5, pp. 868–869. ISSN: 1573-9686. DOI: 10.1007/s10439-023-03172-7. URL: https://doi.org/10.1007/s10439-023-03172-7 (visited on 04/01/2024).

Bodie, Zvi, Alex Kane and Alan J. Marcus (2021). *Investments*. Twelfth edition, International student edition. The McGraw-Hill Education Series in Finance, Insurance, and Real Estate. New York, NY: McGraw-Hill. 966 pp. ISBN: 978-1-260-57115-8.

Boido, Claudio and Antonio Fasano (3rd Nov. 2014). *CAPM with Sentiment: The Efficient Market Hypothesis Spiced Up with Sentiment*. DOI: 10.2139/ssrn.2550669. URL: https://papers.ssrn.com/abstract=2550669 (visited on 06/01/2024). preprint.

Bollen, Johan, Huina Mao and Xiaojun Zeng (1st Mar. 2011). "Twitter Mood Predicts the Stock Market". In: *Journal of Computational Science* 2.1, pp. 1–8. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2010.12.007. URL: https://www.sciencedirect.com/science/article/pii/S187775031100007X (visited on 03/01/2024).

Bondt, Werner F. M. De and Richard Thaler (1985). "Does the Stock Market Overreact?" In: *The Journal of Finance* 40.3, pp. 793–805. ISSN: 0022-1082. DOI: 10.2307/2327804. JSTOR: 2327804. URL: https://www.jstor.org/stable/2327804 (visited on 03/01/2024).

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan and Matthew Richardson (Jan. 2012). "Which News Moves Stock Prices? A Textual Analysis". In.

Bousquet, Olivier, Ulrike von Luxburg and Gunnar Rätsch, eds. (2004). *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003: Revised Lectures*. Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence 3176. Berlin ; New York: Springer. 240 pp. ISBN: 978-3-540-23122-6.

Branch, Ben (1976). "The Predictive Power of Stock Market Indicators". In: *The Journal of Financial and Quantitative Analysis* 11.2, pp. 269–285. ISSN: 0022-1090. DOI: 10.2307/2979054. JSTOR: 2979054. URL: https://www.jstor.org/stable/2979054 (visited on 03/01/2024).

Breiman, Leo (1st Oct. 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 24/12/2023).

Breitung, Christian, Garvin Kruthof and Sebastian Müller (27th Oct. 2023). *Contextualized Sentiment Analysis Using Large Language Models*. DOI: 10.2139/ssrn.4615038. URL: https://papers.ssrn.com/abstract=4615038 (visited on 20/12/2023). preprint.

Brown and Michael Cliff (1st Jan. 2004). "Investor Sentiment and the Near-Term Stock Market". In: *Journal of Empirical Finance* 11.1, pp. 1–27. ISSN: 0927-5398. DOI: 10.1016/j.jempfin.2002.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0927539803000422 (visited on 05/01/2024).

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher

Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei (22nd July 2020). *Language Models Are Few-Shot Learners*. DOI: 10.48550/arXiv.2005.14165. arXiv: 2005.14165 [cs]. URL: http://arxiv.org/abs/2005.14165 (visited on 27/12/2023). preprint.

Carhart, Mark M. (1997). "On Persistence in Mutual Fund Performance". In: *The Journal of Finance* 52.1, pp. 57–82. ISSN: 0022-1082. DOI: 10.2307/2329556. JSTOR: 2329556. URL: https://www.jstor.org/stable/2329556 (visited on 03/01/2024).

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang and Xing Xie (17th Oct. 2023). *A Survey on Evaluation of Large Language Models*. arXiv: 2307.03109 [cs]. URL: http://arxiv.org/abs/2307.03109 (visited on 20/12/2023). preprint.

Colacicchi, Leonardo (1st Jan. 2022). "Comparison and Fine-Tuning of Methods for Financial Sentiment Analysis". In.

Dan Florin, Stanescu and Marius Romascanu (31st Dec. 2021). "Applying Technology Acceptance Model (TAM) to Explore Users' Behavioral Intention to Adopt Wearables Technologies". In.

Desola, Vinicio, Kevin Hanna and Pri Nonis (5th Aug. 2019). *FinBERT: Pre-Trained Model on SEC Filings for Financial Natural Language Tasks*. DOI: 10.13140/RG.2.2.19153.89442.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (24th May 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 22/12/2023). preprint.

Diebold, Francis X., Jinyong Hahn and Anthony S. Tay (Nov. 1999). "Multivariate Density Forecast Evaluation and Calibration In Financial Risk Management: High-Frequency Returns on Foreign Exchange". In: *Review of Economics and Statistics* 81.4, pp. 661–673. ISSN: 0034-6535, 1530-9142. DOI: 10.1162/003465399558526. URL: https://direct.mit.edu/rest/article/81/4/661-673/57171 (visited on 05/01/2024).

Ding, Xiao, Yue Zhang, Ting Liu and Junwen Duan (Oct. 2014). "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2014. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1415–1425. DOI: 10.3115/v1/D14-1148. URL: https://aclanthology.org/D14-1148 (visited on 03/01/2024).

Eleswarapu, Venkat R. and Marc R. Reinganum (2004). "The Predictability of Aggregate Stock Market Returns: Evidence Based on Glamour Stocks". In: *The Journal* *of Business* 77.2, pp. 275–294. ISSN: 0021-9398. DOI: 10.1086/381275. JSTOR: 10.1086/381275. URL: https://www.jstor.org/stable/10.1086/381275 (visited on 03/01/2024).

Ericsson, Johan and Sune Karlsson (Apr. 2004). "Choosing Factors in a Multifactor Asset Pricing Model: A Bayesian Approach". In: *Stockholm School of Economics*.

Esqueda, Omar and Thomas O'Connor (1st Jan. 2023). "The Cost of Equity to Earnings Yield Differential and Dividend Policy". In: *Managerial Finance* ahead-of-print (ahead-of-print). ISSN: 0307-4358. DOI: 10.1108/MF-01-2023-0063. URL: https://doi.org/10.1108/MF-01-2023-0063 (visited on 21/12/2023).

Fama, Eugene F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, pp. 383–417. ISSN: 0022-1082. DOI: 10.2307/2325486. JSTOR: 2325486. URL: https://www.jstor.org/stable/2325486 (visited on 03/01/2024).

Fama, Eugene F. and Kenneth R. French (1992). "The Cross-Section of Expected Stock Returns". In: *The Journal of Finance* 47.2, pp. 427–465. ISSN: 0022-1082. DOI: 10.2307/2329112. JSTOR: 2329112. URL: https://www.jstor.org/stable/2329112 (visited on 03/01/2024).

Fatouros, Georgios, John Soldatos, Kalliopi Kouroumali, Georgios Makridis and Dimosthenis Kyriazis (Dec. 2023). "Transforming Sentiment Analysis in the Financial Domain with ChatGPT". In: *Machine Learning with Applications* 14, p. 100508. ISSN: 26668270. DOI: 10.1016/j.mlwa.2023.100508. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666827023000610 (visited on 18/11/2023).

Ferreira, Fernando G. D. C., Amir H. Gandomi and Rodrigo T. N. Cardoso (2021). "Artificial Intelligence Applied to Stock Market Trading: A Review". In: *IEEE Access* 9, pp. 30898–30917. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3058133. URL: https://ieeexplore.ieee.org/document/9350582 (visited on 03/01/2024).

Gal, Uri (2023). *ChatGPT Is a Data Privacy Nightmare*. The University of Sydney. URL: https://www.sydney.edu.au/news-opinion/news/2023/02/08/chatgpt-is-a-data-privacy-nightmare.html (visited on 27/12/2023).

Gandhmal, Dattatray P. and K. Kumar (1st Nov. 2019). "Systematic Analysis and Review of Stock Market Prediction Techniques". In: *Computer Science Review* 34, p. 100190. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2019.08.001. URL: https://www.sciencedirect.com/science/article/pii/S157401371930084X (visited on 03/01/2024).

Gao, Tianyu (3rd July 2021). *Prompting: Better Ways of Using Language Models for NLP Tasks*. The Gradient. URL: https://thegradient.pub/prompting/ (visited on 27/12/2023).

Gao, Tianyu, Adam Fisch and Danqi Chen (2nd June 2021). *Making Pre-trained Language Models Better Few-shot Learners*. DOI: 10.48550/arXiv.2012.15723. arXiv:

2012.15723 [cs]. URL: http://arxiv.org/abs/2012.15723 (visited on 04/01/2024). preprint.

Genuer, Robin, Jean-Michel Poggi and Christine Tuleau (21st Nov. 2008). *Random Forests: Some Methodological Insights*. arXiv: 0811.3619 [stat]. URL: http://arxiv.org/abs/0811.3619 (visited on 25/12/2023). preprint.

Godsay, Manasee (17th Sept. 2015). "The Process of Sentiment Analysis: A Study". In: *International Journal of Computer Applications* 126.7, pp. 26–30. ISSN: 09758887. DOI: 10.5120/ijca2015906091. URL: http://www.ijcaonline.org/research/volume126/number7/godsay-2015-ijca-906091.pdf (visited on 27/12/2023).

Google (2023). "Introduction to Large Language Models — Machine Learning". In: *Google for Developers* (). URL: https://developers.google.com/machine-learning/resources/intro-llms (visited on 27/12/2023).

Griffith, Eric (2023). *GPT-4 vs. ChatGPT-3.5: What's the Difference?* PCMAG. URL: https://www.pcmag.com/news/the-new-chatgpt-what-you-get-with-gpt-4-vs-gpt-35 (visited on 27/12/2023).

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7. URL: http://link.springer.com/10.1007/978-0-387-84858-7 (visited on 24/12/2023).

Horton, Robert, Maryam Tavakoli, Alexandre Vilcek, Wolfgang M Pauli and Mario E Inchiosa (2021). "Generating and Evaluating Simulated Medical Notes: Getting a Natural Language Generation Model to Give You What You Want". In.

Huang, Yuxuan, Luiz Fernando Capretz and Danny Ho (Dec. 2021). "Machine Learning for Stock Prediction Based on Fundamental Analysis". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01–10. DOI: 10.1109/SSCI50451.2021.9660134. URL: https://ieeexplore.ieee.org/document/9660134 (visited on 03/01/2024).

Insuasti, Jesus, Felipe Roa and Carlos Mario Zapata-Jaramillo (14th Dec. 2023). "Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models". In: *Big Data and Cognitive Computing* 7.4, p. 182. ISSN: 2504-2289. DOI: 10.3390/bdcc7040182. URL: https://www.mdpi.com/2504-2289/7/4/182 (visited on 27/12/2023).

Inthachot, Montri, Veera Boonjing and Sarun Intakosum (15th Nov. 2016). "Artificial Neural Network and Genetic Algorithm Hybrid Intelligence for Predicting Thai Stock Price Index Trend". In: *Computational Intelligence and Neuroscience* 2016, e3045254. ISSN: 1687-5265. DOI: 10.1155/2016/3045254. URL: https://www.hindawi.com/journals/cin/2016/3045254/ (visited on 03/01/2024).

Istiake Sunny, Md. Arif, Mirza Mohd Shahriar Maswood and Abdullah G. Alharbi (Oct. 2020). "Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model". In: *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 87–92. DOI: 10.1109/NILES50944.2020.9257950. URL: https://ieeexplore.ieee.org/abstract/document/9257950 (visited on 04/01/2024).

Jacobs, Ph D., Kenneth N. Levy and David Starer (11th Aug. 1998). *Long-Short Portfolio Management: An Integrated Approach*. URL: https://papers.ssrn.com/abstract=2858554 (visited on 06/01/2024). preprint.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency". In: *The Journal of Finance* 48.1, pp. 65–91. ISSN: 0022-1082. DOI: 10.2307/2328882. JSTOR: 2328882. URL: https://www.jstor.org/stable/2328882 (visited on 03/01/2024).

Jensen, Michael C. (1968). "The Performance of Mutual Funds in the Period 1945–1964". In: *The Journal of Finance* 23.2, pp. 389–416. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1968.tb00815.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1968.tb00815.x (visited on 03/01/2024).

Johnson (24th Sept. 2015). "Multi-Factor Extensions of the Capital Asset Pricing Model: An Empirical Study of the UK Market". Västerås, Sweden: Mälardalen University. URL: https://www.semanticscholar.org/paper/Multi-Factor-Extensions-of-the-Capital-Asset-Model%3A-Johnson/62ad909660feac2cfa9b45dc07021941ec69962f (visited on 05/01/2024).

Jordan, M. I. and T. M. Mitchell (17th July 2015). "Machine Learning: Trends, Perspectives, and Prospects". In: *Science* 349.6245, pp. 255–260. DOI: 10.1126/science.aaa8415. URL: https://www.science.org/doi/full/10.1126/science.aaa8415 (visited on 25/12/2023).

Kearney, Colm and Sha Liu (1st May 2014). "Textual Sentiment in Finance: A Survey of Methods and Models". In: *International Review of Financial Analysis* 33, pp. 171–185. ISSN: 1057-5219. DOI: 10.1016/j.irfa.2014.02.006. URL: https://www.sciencedirect.com/science/article/pii/S1057521914000295 (visited on 20/12/2023).

Kim, Jihwan, Hui-Sang Kim and Sun-Yong Choi (29th Aug. 2023). "Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM". In: *Axioms* 12.9, p. 835. ISSN: 2075-1680. DOI: 10.3390/axioms12090835. URL: https://www.mdpi.com/2075-1680/12/9/835 (visited on 27/12/2023).

Kingsford, Carl and Steven L Salzberg (Sept. 2008). "What Are Decision Trees?" In: *Nature Biotechnology* 26.9, pp. 1011–1013. ISSN: 1087-0156, 1546-1696. DOI: 10.

1038/nbt0908-1011. URL: https://www.nature.com/articles/nbt0908-1011 (visited on 27/12/2023).

Kuhn, Max and Kjell Johnson (2013). *Applied Predictive Modeling*. New York, NY: Springer. ISBN: 978-1-4614-6848-6 978-1-4614-6849-3. DOI: 10.1007/978-1-4614-6849-3. URL: http://link.springer.com/10.1007/978-1-4614-6849-3 (visited on 03/01/2024).

Liu, Bing (2017). "Many Facets of Sentiment Analysis". In: *A Practical Guide to Sentiment Analysis*. Ed. by Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay and Antonio Feraco. Socio-Affective Computing. Cham: Springer International Publishing, pp. 11–39. ISBN: 978-3-319-55394-8. DOI: 10.1007/978-3-319-55394-8_2. URL: https://doi.org/10.1007/978-3-319-55394-8_2 (visited on 03/01/2024).

Lopez-Lira, Alejandro and Yuehua Tang (9th Sept. 2023). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. arXiv: 2304.07619 [cs, q-fin]. URL: http://arxiv.org/abs/2304.07619 (visited on 04/01/2024). preprint.

Madge, Saahil and Professor Swati Bhatt (n.d.). "Predicting Stock Price Direction Using Support Vector Machines". In: ().

Malkiel, Burton G. (Mar. 2003). "The Efficient Market Hypothesis and Its Critics". In: *Journal of Economic Perspectives* 17.1, pp. 59–82. ISSN: 0895-3309. DOI: 10.1257/089533003321164958. URL: https://www.aeaweb.org/articles?id=10.1257/089533003321164958 (visited on 27/12/2023).

Malo, Pekka, Ankur Sinha, Pyry Takala, Pekka Korhonen and Jyrki Wallenius (1st Apr. 2014). "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts". In: *Journal of the American Society for Information Science and Technology*. DOI: 10.1002/asi.23062.

Medhat, Walaa, Ahmed Hassan and Hoda Korashy (1st Dec. 2014). "Sentiment Analysis Algorithms and Applications: A Survey". In: *Ain Shams Engineering Journal* 5.4, pp. 1093–1113. ISSN: 2090-4479. DOI: 10.1016/j.asej.2014.04.011. URL: https://www.sciencedirect.com/science/article/pii/S2090447914000550 (visited on 03/01/2024).

Meskó, Bertalan (4th Oct. 2023). "Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial". In: *Journal of Medical Internet Research* 25.1, e50638. DOI: 10.2196/50638. URL: https://www.jmir.org/2023/1/e50638 (visited on 27/12/2023).

Microsoft (6th Dec. 2023). *Metodtips: Hyperparameterjustering med Hyperopt – Azure Databricks*. URL: https://learn.microsoft.com/sv-se/azure/databricks/machine-learning/automl-hyperparam-tuning/hyperopt-best-practices (visited on 08/01/2024).

Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev and Dimitar Trajanov (2020). "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers". In: *IEEE Access* 8, pp. 131662–131682. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3009626. URL: https://ieeexplore.ieee.org/document/9142175 (visited on 03/01/2024).

Moin, Muhammad Shehzad (2023). *Corporate Sustainability and Working Capital : A Panel Data Analysis of the Relationship in Swedish-listed Firms*. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-211067 (visited on 21/12/2023).

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike and Ryan Lowe (4th Mar. 2022). *Training Language Models to Follow Instructions with Human Feedback*. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155 [cs]. URL: http://arxiv.org/abs/2203.02155 (visited on 27/12/2023). preprint.

Park, Cheol-Ho and Scott H. Irwin (2007). "What Do We Know About the Profitability of Technical Analysis?" In: *Journal of Economic Surveys* 21.4, pp. 786–826. ISSN: 1467-6419. DOI: 10.1111/j.1467-6419.2007.00519.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2007.00519.x (visited on 03/01/2024).

Pavlyshenko, Bohdan M. (11th Sept. 2023). *Financial News Analytics Using Fine-Tuned Llama 2 GPT Model*. arXiv: 2308.13032 [cs]. URL: http://arxiv.org/abs/2308.13032 (visited on 22/12/2023). preprint.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga and Diyi Yang (Dec. 2023). "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 1339–1384. DOI: 10.18653/v1/2023.emnlp-main.85. URL: https://aclanthology.org/2023.emnlp-main.85 (visited on 04/01/2024).

Reinganum, Marc R. (1981). "A New Empirical Perspective on the CAPM". In: *The Journal of Financial and Quantitative Analysis* 16.4, pp. 439–462. ISSN: 0022-1090. DOI: 10.2307/2330365. JSTOR: 2330365. URL: https://www.jstor.org/stable/2330365 (visited on 05/01/2024).

Rouf, Nusrat, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich and Hee-Cheol Kim (Jan. 2021). "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions". In: *Electronics* 10.21 (21), p. 2717. ISSN: 2079-9292. DOI: 10.3390/electronics10212717. URL: https://www.mdpi.com/2079-9292/10/21/2717 (visited on 03/01/2024).

Scikit (2023). *Comparing Randomized Search and Grid Search for Hyperparameter Estimation*. scikit-learn. URL: https://scikit-learn/stable/auto_examples/model_selection/plot_randomized_search.html (visited on 05/01/2024).

Shapiro, Adam Hale, Moritz Sudhof and Daniel Wilson (1st Mar. 2020). *Measuring News Sentiment*. 2017-01. Federal Reserve Bank of San Francisco. URL: https://www.frbsf.org/economic-research/publications/working-papers/2017/01/ (visited on 03/01/2024).

Shiller, Robert J. (2000). *Irrational Exuberance*. Princeton: Princeton Univ. Press. 296 pp. ISBN: 978-0-691-05062-1 978-0-691-08915-7.

– (2003). "From Efficient Markets Theory to Behavioral Finance". In: *The Journal of Economic Perspectives* 17.1, pp. 83–104. ISSN: 0895-3309. JSTOR: 3216841. URL: https://www.jstor.org/stable/3216841 (visited on 03/01/2024).

Shmueli, Galit (Aug. 2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/10-STS330. URL: https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full (visited on 03/01/2024).

Soong, Gim Hoy and Chye Cheah Tan (Nov. 2021). "Sentiment Analysis on 10-K Financial Reports Using Machine Learning Approaches". In: *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*. 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), pp. 124–129. DOI: 10.1109/ICSET53708.2021.9612552. URL: https://ieeexplore.ieee.org/abstract/document/9612552 (visited on 27/12/2023).

Soroka, Stuart, Patrick Fournier and Lilach Nir (2023). *Cross-National Evidence of a Negativity Bias in Psychophysiological Reactions to News*. In collab. with Susan Fiske. URL: https://www.pnas.org/doi/10.1073/pnas.1908369116 (visited on 22/12/2023).

Stattman, Dennis (1980). "Book Values and Expected Stock Returns," Chicago, IL: University of Chicago.

Stone, Philip, Dexter Dunphy, Marshall Smith and Daniel Ogilvie (1st Jan. 1966). *The General Inquirer: A Computer Approach to Content Analysis*. Vol. 4. DOI: 10.2307/1161774.

Tashman, Leonard J. (1st Oct. 2000). "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review". In: *International Journal of Forecasting*. The M3- Competition 16.4, pp. 437–450. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(00)00065-0. URL: https://www.sciencedirect.com/science/article/pii/S0169207000000650 (visited on 06/01/2024).

Tetlock, Paul C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *The Journal of Finance* 62.3, pp. 1139–1168. ISSN: 1540-

6261. DOI: 10.1111/j.1540-6261.2007.01232.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2007.01232.x (visited on 22/12/2023).

Ticknor, Jonathan L. (15th Oct. 2013). “A Bayesian Regularized Artificial Neural Network for Stock Market Forecasting”. In: *Expert Systems with Applications* 40.14, pp. 5501–5506. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2013.04.013. URL: https://www.sciencedirect.com/science/article/pii/S0957417413002509 (visited on 03/01/2024).

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov and Thomas Scialom (19th July 2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs]. URL: http://arxiv.org/abs/2307.09288 (visited on 21/12/2023). preprint.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 04/01/2024).

Verma, Rahul and Priti Verma (1st July 2007). “Noise Trading and Stock Market Volatility”. In: *Journal of Multinational Financial Management* 17.3, pp. 231–243. ISSN: 1042-444X. DOI: 10.1016/j.mulfin.2006.10.003. URL: https://www.sciencedirect.com/science/article/pii/S1042444X06000806 (visited on 06/01/2024).

Wang, Jin, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang (Aug. 2016). “Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2016. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 225–230. DOI: 10.18653/v1/P16-2037. URL: https://aclanthology.org/P16-2037 (visited on 04/01/2024).

Xing, Frank Z., Erik Cambria and Roy E. Welsch (Nov. 2018a). "Intelligent Asset Allocation via Market Sentiment Views". In: *IEEE Computational Intelligence Magazine* 13.4, pp. 25–34. ISSN: 1556-6048. DOI: 10.1109/MCI.2018.2866727. URL: https://ieeexplore.ieee.org/document/8492384 (visited on 03/01/2024).

– (1st June 2018b). "Natural Language Based Financial Forecasting: A Survey". In: *Artificial Intelligence Review* 50.1, pp. 49–73. ISSN: 1573-7462. DOI: 10.1007/s10462-017-9588-9. URL: https://doi.org/10.1007/s10462-017-9588-9 (visited on 21/12/2023).

Xu, Yun and Royston Goodacre (July 2018). "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning". In: *Journal of Analysis and Testing* 2.3, pp. 249–262. ISSN: 2096-241X, 2509-4696. DOI: 10.1007/s41664-018-0068-2. URL: http://link.springer.com/10.1007/s41664-018-0068-2 (visited on 24/12/2023).

Ying, Xue (Feb. 2019). "An Overview of Overfitting and Its Solutions". In: *Journal of Physics: Conference Series* 1168.2, p. 022022. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1168/2/022022. URL: https://dx.doi.org/10.1088/1742-6596/1168/2/022022 (visited on 24/12/2023).

Zaremba, Adam and Ender Demir (13th Jan. 2023). *ChatGPT: Unlocking the Future of NLP in Finance*. DOI: 10.2139/ssrn.4323643. URL: https://papers.ssrn.com/abstract=4323643 (visited on 04/01/2024). preprint.

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan and Jimmy Ba (10th Mar. 2023). *Large Language Models Are Human-Level Prompt Engineers*. arXiv: 2211.01910 [cs]. URL: http://arxiv.org/abs/2211.01910 (visited on 27/12/2023). preprint.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano and Geoffrey Irving (8th Jan. 2020). *Fine-Tuning Language Models from Human Preferences*. DOI: 10.48550/arXiv.1909.08593. arXiv: 1909.08593 [cs, stat]. URL: http://arxiv.org/abs/1909.08593 (visited on 27/12/2023). preprint.

Appendix

Table 1: Summary of Python libraries used

Library	**Description**
pandas	Data manipulation and analysis
numpy	Numerical operations and array handling
os	Operating system interfaces and file handling
langchain.*	Libraries related to language chain operations
json	JSON file parsing and handling
statsmodels.api	Statistical models, including OLS
matplotlib.pyplot	Plotting and visualization
re	Regular expression operations
platform	Access to underlying platform's identifying data
sklearn.*	Machine learning models and evaluation tools
seaborn	Advanced statistical data visualization

Table 2: OMX Stockholm 30 index, leavers and joiners - retrieved from Refinitiv Eikon

Status	Company	Ticker	Date
+	Evolution AB	EVOG	2021-01-04
-	SSAB	SSABa	2021-01-04
-	Securitas	SECUb	2021-07-01
+	Sinch	SINCH	2021-07-01
+	Samhall Norden	SBBb	2022-07-01
-	Skanska AB	SKAb	2022-07-01
-	Swedish Match	SWMA.ST^A23 (expired)	2022-11-17
+	Nibe Industrier	NIBEb.ST	2023-01-02

Table 3: Expansion of company names for all tickers in OMXS30

Ticker	**Company name**
ASSA-B	Assa Abloy AB
ESSITY-B	Essity AB
TELIA	Telia Company AB
INVE-B	Investor AB
SINCH	Sinch AB
GETI-B	Getinge AB
SWED-A	Swedbank AB
SKF-B	Aktiebolaget SKF
TEL2-B	Tele2 AB
SCA-B	Svenska Cellulosa Aktiebolaget SCA
NDA-SE	Nordea Bank Abp
ATCO-A	Atlas Copco Aktiebolag
AZN	Astra Zeneca plc
VOLV-B	Aktiebolaget Volvo
SEB-A	Skandinaviska Enskilda Banken AB (SEB)
SHB-A	Svenska Handelsbanken AB
ALIV-SDB	Autoliv AB
KINV-B	Kinnevik AB
ABB	ABB Ltd
ELUX-B	Aktiebolaget Electrolux
BOL	Boliden AB
HM-B	H & M Hennes & Mauritz AB (H&M)
NIBE-B	Nibe Industrier AB
EVO	Evolution AB
ALFA	Alfa Laval AB
SAND	Sandvik Aktiebolag
SBB-B	Samhällsbyggnadbolaget i Norden AB
HEXA-B	Hexagon AB
ERIC-B	Telefonaktiebolaget LM Ericsson

Table 4: Key parameters used in Llama, along with their default values

Parameter	Function	Default value
temperature	Controls the randomness of the output	0.8
max_tokens	Maximal amount of output tokens	512
top_p	Balances uniqueness	0.95
verbose	Used for debugging	True

Listing 1: Original prompt

```
"GPT-P4": {
      "messages": [
          {
              "role": "system",
              "content": "Act as
                  an expert at
                  forex trading
                  holding {ticker}"
          },
          {
              "role": "assistant",
              "content": "Based
                  only on the
                  headline
                  '{headline}',
                  will you buy,
                  sell or hold
                  {ticker} in the
                  short term?"
          },
          {
              "role": "user",
              "content": "Answer
                  in one token:
                  positive for buy,
                  negative for
                  sell, or neutral
                  for hold position"
          }
      ],
      "max_tokens": 1,
      "top_p": 1
  },
```

Listing 2: Transformed prompt

```
"""
        [INST] <<SYS>>
            Act as an expert in
                stock trading
                holding {ticker}.
            Answer in one token:
                positive for buy,
                negative for sell,
                or neutral for hold
                position.
        <</SYS>>
            Based only on the
                headline '{text}',
                will you buy, sell
                or hold {ticker} in
                the short term?
        [/INST]
"""
```

Figure 1: Transformation of GPT-P4 prompt into Llama syntax