“Enhancing portfolio optimization with multi-LLM sentiment aggregation: A Black-Litterman integration approach”

AUTHORS

Lamukanyani Alson Mantshimuli
John Weirstrass Muteba Mwamba

ARTICLE INFO	Lamukanyani Alson Mantshimuli and John Weirstrass Muteba Mwamba (2025). Enhancing portfolio optimization with multi-LLM sentiment aggregation: A Black-Litterman integration approach. *Investment Management and Financial Innovations*, *22*(3), 213-226. doi:10.21511/imfi.22(3).2025.16
DOI	http://dx.doi.org/10.21511/imfi.22(3).2025.16
RELEASED ON	Thursday, 14 August 2025
RECEIVED ON	Tuesday, 11 March 2025
ACCEPTED ON	Thursday, 10 July 2025
LICENSE	(cc) BY This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
30	2	7

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

Lamukanyani Alson Mantshimuli (South Africa),
John Weirstrass Muteba Mwamba (South Africa)

ENHANCING PORTFOLIO OPTIMIZATION WITH MULTI-LLM SENTIMENT AGGREGATION: A BLACK-LITTERMAN INTEGRATION APPROACH

Type of the article: Research Article

Received on: 11th of March, 2025
Accepted on: 10th of July, 2025
Published on: 14th of August, 2025

Lamukanyani Alson Mantshimuli, PhD Candidate, Department of Finance and Investment, College of Business and Economics, University of Johannesburg, South Africa. (Corresponding author)

John Weirstrass Muteba Mwamba, Visiting Associate Professor, School of Economics, College of Business and Economics, University of Johannesburg, South Africa.

Abstract

Sentiment analysis of financial text data plays a crucial role in investment decision-making, yet existing approaches often rely on single-model sentiment scores that may suffer from biases or hallucinations. This study aims to enhance portfolio optimization by integrating sentiment signals from multiple Large Language Models (LLMs) into the Black-Litterman framework. The proposed method aggregates sentiment scores from three finance-domain fine-tuned LLMs using a Long Short-Term Memory network, which captures non-linear relationships and temporal dependencies to produce a robust Meta-LLM sentiment score. This score is then incorporated into the Black-Litterman model as investor views to derive optimal portfolio weights. The methodology is tested on a portfolio of S&P 500 stocks. The results show that the proposed approach significantly improves portfolio performance, achieving an annualized return of 31.22%, compared to 24.57% for the market capital-weighted portfolio. Additionally, the model attains a Sharpe Ratio of 3.02, an Omega Ratio of 2.48, and a Jensen's Alpha of 1.95%, outperforming both the benchmark portfolios and portfolios based on single-LLM sentiment. The findings demonstrate that aggregating sentiment from multiple LLMs enhances risk-adjusted returns while mitigating model-specific limitations. Future research could explore the integration of LLMs with different architectures to further refine sentiment-aware portfolio strategies.

Keywords large language models, sentiment analysis, long short-term memory, portfolio optimization, Black-Litterman model, financial text data

JEL Classification G11

INTRODUCTION

Large language models (LLMs) for portfolio optimization represent a significant step in leveraging advancements in artificial intelligence and machine learning for asset allocation and security selection decisions. With their ability to process and interpret vast amounts of financial text data, including news articles and social media posts, LLMs have emerged as powerful tools for capturing market sentiment. The integration of sentiment analysis into quantitative portfolio optimization, however, remains a critical challenge in modern finance. While Large Language Models (LLMs) have demonstrated exceptional capabilities in extracting sentiment from financial text, their application in asset allocation is often limited by two key issues: (1) reliance on single-model sentiment scores, which are prone to biases or hallucinations, and (2) the inability of traditional portfolio frameworks to dynamically incorporate nuanced, multi-source sentiment signals. These limitations hinder the development of robust, sentiment-aware investment strategies that can adapt to rapidly changing market conditions.

Conflict of interest statement:
Author(s) reported no conflict of interest

http://dx.doi.org/10.21511/imfi.22(3).2025.16

The research problem at the core of this study is the lack of a systematic approach to aggregate and leverage sentiment signals from multiple LLMs within established portfolio optimization frameworks. Additionally, existing methods either treat LLM-generated sentiment as static inputs or fail to account for the temporal and contextual dependencies between disparate sentiment signals. This gap is particularly consequential for the Black-Litterman model (Black & Litterman, 1990), which relies on accurate views to adjust portfolio weights but traditionally derives these views from human judgment, not from aggregated, machine-learned sentiment. This study addresses these challenges by proposing a Meta-LLM sentiment aggregation framework that unifies outputs from three finance-specific LLMs: FinBERT Prosus AI (Araci, 2019), FinBERT Yiyanghkust (Huang et al., 2021), and Distil RoBERTa (Romero, 2024).

The integration of multiple LLMs enhances sentiment analysis robustness by leveraging their complementary strengths derived from distinct training corpora and fine-tuning objectives. This multi-model approach captures diverse linguistic perspectives while mitigating individual model biases. These outputs are aggregated through an LSTM network, which effectively models temporal dependencies in sentiment evolution, to create dynamic sentiment views for the Black-Litterman framework, effectively bridging high-frequency textual analysis with traditional portfolio optimization. This synthesis enables more adaptive asset allocation that responds to evolving market sentiment while maintaining the theoretical rigor of established financial models.

1. THEORETICAL BASIS

Over the past few years, the usage of LLMs has increased exponentially, with the introduction of popular chatbots such as ChatGPT further growing the popularity of LLMs. The study focuses on the usage of LLMs for sentiment analysis in the finance domain. The field of sentiment analysis has undergone a significant shift in recent years, transitioning from traditional Natural Language Processing (NLP) techniques to the adoption of Large Language Models (LLMs). Early approaches to sentiment analysis relied on rule-based systems, machine learning algorithms, and NLP techniques, such as those proposed by Pang and Lee (2008) and Liu (2012). However, these methods often struggled with nuances of language, context, and ambiguity. The arrival of LLMs has transformed sentiment analysis by enabling the capture of complex linguistic patterns and contextual relationships. LLMs have been shown to outperform traditional NLP approaches in sentiment analysis tasks, as demonstrated by studies such as Sun et al. (2019) and Zhang et al. (2020). The ability of LLMs to learn from vast amounts of text data has made them an attractive choice for general sentiment analysis applications.

While large language models (LLMs) outperform traditional NLP methods in general sentiment analysis, their direct application to financial texts often yields suboptimal results due to misalignment between pre-training objectives and domain-specific sentiment labeling (Zhang et al., 2023). This gap stems from financial language's specialized terminology and nuanced context, which general-purpose models like BERT struggle to capture effectively. However, such models can be adapted for finance through targeted fine-tuning as demonstrated by Araci (2019) with FinBERT, a BERT-based model fine-tuned on financial corpora that surpasses both generic LLMs and traditional machine learning techniques in sentiment analysis. Similarly, Lou et al. (2024) advanced this approach by applying supervised fine-tuning and few-shot prompting to adapt Llama-7B (Touvron et al., 2023) for financial texts, achieving state-of-the-art accuracy. These studies collectively highlight that domain adaptation, whether via retrieval augmentation (Zhang et al., 2023), fine-tuning (Araci, 2019), or few-shot learning (Lou et al., 2024), is critical for bridging the performance gap in financial sentiment analysis.

While employing these domain adaptation strategies enhances sentiment analysis accuracy, the standalone utility of LLM-computed sentiment remains limited. Zhao et al. (2024) emphasize that LLMs sentiment output achieve maximal impact when integrated with traditional quantitative models, a principle central to the study's design. Augmenting LLM sentiment analysis with traditional quantitative methods in the finance domain offers a promising way to capture both market sentiment and financial metrics, enhancing stock market prediction and decision-making in portfolio optimization and asset allocation. Recent work by Kuruvilla and Mythily (2025) demonstrates the transformative potential of such hybrid approaches, showing that integrating LLM-processed sentiment with temporal forecasting models can significantly improve risk-adjusted returns. Furthermore, Kirtac and Germano (2025) have recently demonstrated the effectiveness of hybrid approaches through their Sentiment-Augmented PPO (SAPPO) model, which integrates LLaMA 3.3-derived sentiment scores with reinforcement learning to achieve superior Sharpe ratios and reduced portfolio drawdowns.

Although these studies demonstrate improved performance over traditional quantitative methods, the potential of applying LLM-based sentiment analysis in portfolio optimization remains under-explored, with limited integration of these models into conventional financial models that typically rely on historical data and technical indicators. By incorporating LLM sentiment scores, portfolios could benefit from more nuanced and real-time insights, potentially improving risk-adjusted returns. Colasanto et al. (2022) pioneered the integration of LLM-derived sentiment in portfolio optimization by incorporating FinBERT's outputs into the Black-Litterman framework. Their results demonstrated a statistically significant improvement in risk-adjusted returns, with sentiment-augmented portfolios achieving a Sharpe ratio of 1.14 compared to 1.07 for traditional portfolios. Building on this work, the study expands the scope by leveraging sentiment scores from three distinct LLMs in the portfolio optimization process. By strategically aggregating sentiment scores across multiple finance-specific LLMs, three well-documented limitations of single-model approaches are addressed: (1) dataset-specific biases in training corpora (Ranjan et al., 2024), (2) hallucination tendencies in financial sentiment analysis (Kang & Liu, 2023), and (3) narrow domain coverage (Shah et al., 2024).

The ensemble method builds on established findings that combining multiple LLMs improves robustness through cross-model validation (Mao et al., 2025) and complementary specialization. This approach particularly mitigates the volatility found in single-LLM sentiment outputs, while achieving the consensus benefits of using multi-LLMs for financial sentiment analysis as demonstrated by Zhang and Shafiq (2024). The aggregation methodology itself plays a pivotal role in harnessing the benefits of using multi-LLMs instead of a single LLM. Prior approaches have relied on other machine learning techniques such as gradient boosting (Pathuri et al., 2020) and support vector machines (de Kok et al., 2018), but recent literature demonstrates superior performance from LSTM-based meta-models. Specifically, LSTMs capture temporal dependencies in evolving market sentiment regimes (Wang et al., 2018) and model non-linear relationships between conflicting outputs more effectively than linear methods (Bukhari et al., 2020), thereby improving their performance.

Building on the reviewed literature, this study aims to develop an innovative portfolio optimization framework that integrates multi-LLM sentiment analysis with LSTM-based temporal aggregation and Black-Litterman optimization. Each model brings distinct strengths: FinBERT Prosus AI is pre-trained on financial news and fine-tuned for financial sentiment analysis; FinBERT Yiyanghkust uses researcher-labelled sentences from analyst reports for fine-tuning; and DistilRoBERTa is pre-trained using masked language modelling. The choice of these for sentiment analysis in this study is grounded in their proven effectiveness (see Dmonte et al. (2024), for example), open-source availability, and demonstrated superior performance in financial sentiment prediction tasks (Xie et al., 2024; Lefort et al., 2024).

The study specifically addresses three critical limitations in current approaches: (1) single-model bias and hallucinations in financial sentiment analysis, (2) over-reliance on historical data in traditional portfolio methods, and (3) the lack of dynamic sentiment integration in asset allocation decisions. Through this integrated approach, the study seeks to demonstrate superior performance via enhanced risk-adjusted returns (measured by Sharpe and Omega Ratios), improved stability of sentiment signals, and more effective capture of emerging market trends, ultimately advancing the integration of AI-driven sentiment analysis with established portfolio theory. Based on these objectives, the following research hypotheses are formulated to fill the research gap:

H01$_{Null}$: *Portfolios incorporating Meta-LLM sentiment scores do not achieve superior absolute and risk-adjusted returns compared to traditional benchmark portfolios.*

H01$_{Alternative}$: *Portfolios incorporating Meta-LLM sentiment scores do achieve superior absolute and risk-adjusted returns compared to traditional benchmark portfolios.*

H02$_{Null}$: *The LSTM-based aggregation method will not significantly reduce sentiment volatility (and hence portfolio volatility) relative to individual LLM outputs.*

H02$_{Alternative}$: *The LSTM-based aggregation method will significantly reduce sentiment volatility (and hence portfolio volatility) relative to individual LLM outputs.*

H03$_{Null}$: *The LSTM aggregation method does not produce sentiment scores with higher accuracy compared to alternative machine learning methods when integrating multi-LLM outputs.*

H03$_{Alternative}$: *The LSTM aggregation method does produce sentiment scores with higher accuracy compared to alternative machine learning methods when integrating multi-LLM outputs.*

2. METHODOLOGY

The approach involves calculating sentiment scores for individual stocks mentioned in the Financial News and Stock Price Integration Dataset (FNSPID) (Dong et al., 2024), which serves as the foundation for portfolio optimization. Three Large Language Models (LLMs), fine-tuned for the finance domain, are employed to compute sentiment scores. These scores are then aggregated into a Meta-LLM sentiment score using a trained Long Short-Term Memory (LSTM) recurrent neural network. The three LLMs used in the study have been fine-tuned for the finance domain, making them appropriate for this application. The Meta-LLM sentiment scores derived from these fine-tuned LLMs are integrated into the Black-Litterman portfolio optimization framework by adjusting the absolute prior views on the portfolio assets to reflect the sentiment scores. The portfolio is rebalanced monthly, with the average Meta-LLM sentiment score for each stock considered at each rebalancing date.

The FNSPID Data Set contains news articles summarizing market developments across asset classes, including equities. For sentiment analysis, theses financial text data are firstly pre-processed by: (1) extracting sentences referencing S&P 500 constituents (investment universe) through a two-stage approach using Python's NLTK sentence tokenizer followed by regular expression matching; (2) applying case-insensitive RegEx patterns to identify whole-word matches of company names and ticker symbols; and (3) filtering matched sentences for downstream LLM-based sentiment scoring. This preprocessing ensures accurate entity detection while maintaining textual context for reliable sentiment extraction.

Given the stock-specific extracted sentences, sentiment scores are derived using each of the 3 LLMs based on the Bidirectional Encoder Representations from Transformers (BERT) architecture and fine-tuned using financial text data: FinBERT Prosus AI, FinBERT Yiyanghkust, and Distil RoBERTa. The Transformer Library in Python was used to load the LLMs, which are used to calculate sentiment scores. The transformer model follows the structure presented in Figure 1 using stacked self-attention and connected layers for both the encoder (left half) and the decoder (right half).

At its core, the transformer is designed to handle sequential data and uses a stack of identical layers, each with two main components: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network, as shown in Figure 1. The self-attention mechanism allows the model to weigh the importance of different words in a sequence, irrespective of their distance

http://dx.doi.org/10.21511/imfi.22(3).2025.16

Source: Vaswani (2017).

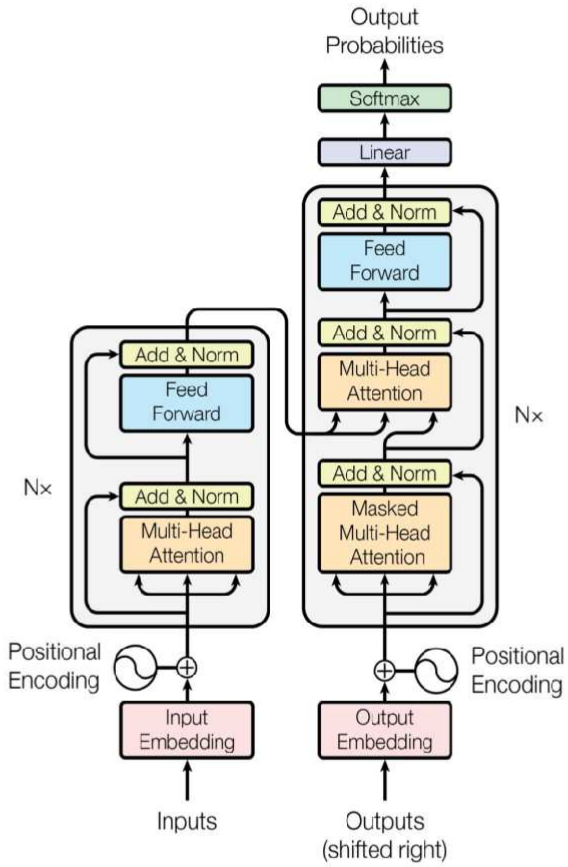

Figure 1. Transformer model architecture

from each other, by creating "attention scores" for each word or token pair. For each token, attention scores are computed based on its relationships with other tokens. Given query, key, and value matrices Q, K, V;

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right), \tag{1}$$

where $Q = X'W^Q$, $K = X'W^K$, $V = X'W^V$ are the query, key, and value matrices, d_k is the dimension of the key vectors, and W^Q, W^K, W^V are learned weight matrices. $X' = X + P$ is the input embeddings combined with positional encodings, i.e., $X = \{x_1,x_2,...,x_n\}$ is the token embeddings for token i and P is the positional encodings. For position i, the encoding is

$$PE_{(i,2k)} = \sin\left(\frac{i}{10000^{2k/d}}\right), \tag{2}$$

$$PE_{(i,2k+1)} = \cos\left(i/10000^{2k/d}\right), \tag{3}$$

where k is the dimension index, and d is the dimensionality of the embedding vector or the model dimension. Multi-head attention allows the model to focus on different parts of the input simultaneously, enhancing its ability to capture various aspects of context.

$$\begin{gathered} MultiHead(Q,K,V) \\ = Concat(head_1,...,head_h)W^O, \end{gathered} \tag{4}$$

where $head_i = Attention\ (Q_i, K_i, V_i)$, and W^O is the output projection matrix. To maintain word order information, the transformer adds positional encodings to each input embedding. This architecture is efficient and parallelizable (computations across the entire input sequence can happen simultaneously), making it faster than recurrent models for long sequences. In the encoder-decoder design of the transformer, the encoder processes the input sequence and produces a set of representations, while the decoder uses these representations, along with the target sequence, to generate the output. Each layer includes residual connections and layer normalization, which help with training stability and enable deeper networks.

In the final layers, the model uses a classification head that transforms the contextualized embeddings into sentiment scores. For each input sentence, the model outputs probabilities corresponding to predefined sentiment classes (e.g., positive, neutral, negative). This classification head, added on top of the final encoder layers, allows the model to perform direct sentiment classification. This process is similar for all 3 fine-tuned LLMs that are used in the analysis. The LLMs' output scores are standardized to a [–1, +1] scale, where +1 indicates strong positive sentiment, and –1 indicates strong negative sentiment. Notably, the sentiment scores from different LLMs often diverge, with some models producing opposite sentiments, e.g., one model yielding a negative score while others yield positive scores on the same set of texts or sentences.

The divergence in sentiment scores across the three LLMs stems from model-specific limitations and hallucination tendencies, where outputs may contradict input data. This variability underscores the advantage of multi-LLM sentiment analysis. An LSTM network is implemented to aggregate these scores into a robust Meta-LLM measure, selected after comparative testing against other machine learning aggregation models (Random Forest, Gradient Boosting, Support Vector Machines (SVMs)) demonstrated its superior performance. The LSTM's gated architecture, comprising forget, input, and output gates, specifically addresses key challenges in sentiment aggregation: temporal processing (handles the sequential nature of sentiment scores as time-series data), dependency modelling (captures both short and long-term relationships between LLM outputs), and gradient control (solves the vanishing gradient problem through regulated information flow).

Each time step processes individual LLM scores, with gates dynamically updating the cell state to compute the final Meta-LLM sentiment. This architecture proves particularly effective for financial applications where sequential patterns in sentiment significantly impact asset prices.

As depicted in Figure 2, the LSTM network is designed with a cell state and multiple gates that work together to process time series data. The cell state acts as a conduit for transmitting relevant information throughout the sequence, while the gates, which utilize sigmoid functions to filter information, selectively retain or discard information. This unique architecture enables LSTMs to capture long-term dependencies in the data, a ca-

Source: Benjamin and Matthew (2025).

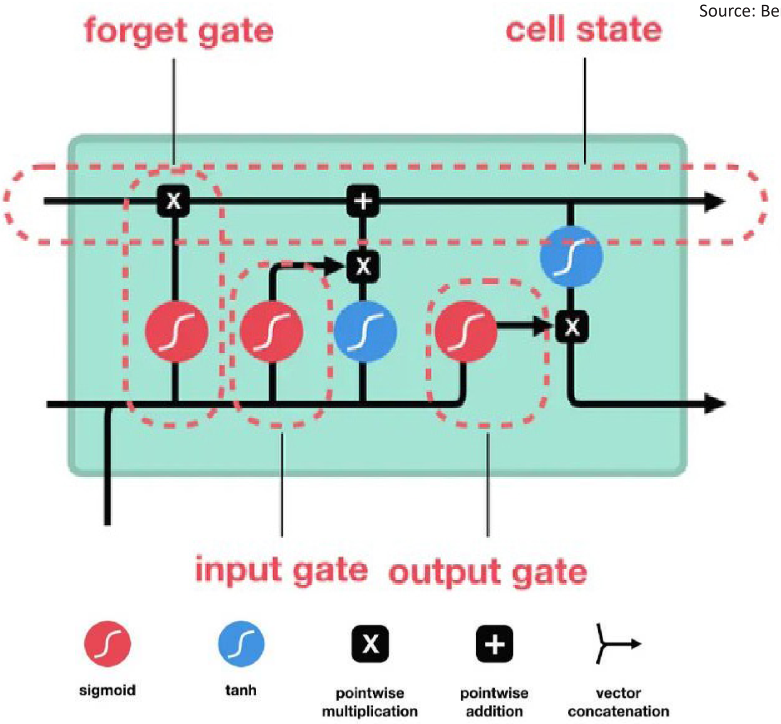

Figure 2. LSTM cell and its operations

http://dx.doi.org/10.21511/imfi.22(3).2025.16

pability that distinguishes them from traditional recurrent neural networks. The process can be summarized by the following equations for each cell or unit:

$$h_t = o_t \odot \tanh(c_t), \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t, \tag{6}$$

$$\widetilde{c}_t = \tanh\left(w_c \left[h_{t-1}, x_t\right] + b_c\right), \tag{7}$$

$$f_t = \sigma\left(w_f \left[h_{t-1}, x_t\right] + b_f\right), \tag{8}$$

$$i_t = \sigma\left(w_i \left[h_{t-1}, x_t\right] + b_i\right), \tag{9}$$

$$o_t = \sigma\left(w_o \left[h_{t-1}, x_t\right] + b_o\right), \tag{10}$$

where f_t, i_t and o_t are the forget, input, and output gates, respectively (at a time t), $\widetilde{c}_t, c_t, h_t$ represent candidate cell state, cell state and the hidden state respectively, $x_t = [s_{i,t}]$ is the input (individual LLM sentiment scores, $s_{i,t}$ where $i = 1,2,3$), h_{t-1} is the hidden state at the previous step which provides the LSTM with memory of previous states, h_t is the hidden state at the current time step (used as the final output for this time step), w_f, w_i, w_c and w_o are weight matrices applied to the concatenation of h_{t-1} and x_t for each respective gate (forget, input, cell and output), b_f, b_i, b_c, and b_o are biases for the respective gates (forget, input, cell and output) which the LSTM learns during training, σ is the sigmoid activation function which outputs values between 0 and 1 to control the flow of information, and *tahn* is the hyperbolic tangent activation function which scales values to keep them between –1 and 1. This is used to scale the candidate cell state and the final hidden-state output.

The final hidden state can then be used as the Meta-LLM sentiment score for the given input sequence, as it captures the aggregated sentiment across all individual LLM scores. The LSTM-generated Meta-LLM sentiment scores are integrated into the Black-Litterman framework by dynamically adjusting absolute return views at each monthly rebalancing date. Using Python's PyPortfolioOpt library, these sentiment-informed views modify the equilibrium returns, allowing the portfolio optimization to incorporate real-time, AI-driven market perspectives while maintaining the model's theoretical rigor.

The Black-Litterman model can be considered a standard model for incorporating views when constructing portfolios. The model combines subjective views of investors (views about the expected return of assets in a portfolio) with the market equilibrium vector of expected returns (prior distribution) to produce a new estimate of expected returns for the portfolio (posterior). The meta-model sentiment scores are used to create views about the expected returns of the assets in a portfolio (absolute return views). Equation 3 shows the formula used to calculate the newly estimated expected returns, which include return views, ($E(R)$).

$$E(R) = \mu = \left[(\tau\Sigma)^{-1} + P^{\top}\Omega^{-1}P\right]^{-1} \times \left[(\tau\Sigma)^{-1}\Pi + P^{\top}\Omega^{-1}Q\right], \tag{11}$$

where τ is a scalar that measures uncertainty or confidence in the prior estimate of returns, Σ is the covariance matrix of excess returns[1] ($N \times N$ matrix, where N is the number of assets in the portfolio), P is the view selection matrix (identifies assets with views) ($K \times N$ matrix, where K is the number of views), Ω is the covariance matrix of error terms from the views and represents uncertainty on each view ($K \times K$ matrix), $\Pi = \delta \sum x_{eq}$ is the implied equilibrium return vector ($N \times 1$ matrix) derived using reverse optimization (δ is the risk aversion coefficient and x_{eq} is the equilibrium weight) and, P is the view vector ($N \times 1$ matrix) which is informed by the Meta-LLM model sentiment scores.

Expected returns calculated by this formula are used by a mean-variance optimizer to find the Black-Litterman optimal portfolio weights at each rebalancing date. The view vector Q will be the absolute return views on assets in the portfolio. The predicted share price[2] is used to compute absolute return views, $Q = [q_1, q_2 ..., q_n]$. The view on a particular asset i, represented by $q_{i,}$ can be calculated as follows;

[1] Excess returns over the domestic short rate or the one-period risk-free rate. This is assumed to be 0 for the purpose of this study.

[2] The study employs an LSTM network to forecast stock prices for portfolio assets, using two key inputs: (1) Meta-LLM sentiment scores (aggregated from three LLMs) and (2) historical price data. The model outputs next-period price predictions, enabling sentiment-informed portfolio decisions.

$$q_i = \ln\left(\frac{S_{T,i}}{S_{0,i}}\right), \tag{12}$$

where $S_{T,i}$ is the predicted price for asset i for the month at the rebalancing date, while $S_{0,i}$ is the current share price for the asset. Expected returns calculated using formula (3) are then used by the mean-variance optimizer to find optimal portfolio weights through maximizing the Sharpe ratio;

$$\max_x \frac{\mu^T x - r_f}{\left(x^T \Sigma x\right)^{1/2}}, \quad s.t. x^T 1 = 1, \tag{13}$$

where x_i is the optimal portfolio weight for asset *I*, and r_f is the risk-free rate.

The optimal portfolio weights are then used to construct the LSTM Meta-LLM Black-Litterman portfolio. The performance of this portfolio is measured and compared against a benchmark portfolio using portfolio performance measures such as the Sharpe Ratio, Omega Ratio, and Jensen's Alpha.

3. RESULTS & DISCUSSION

3.1. Data

This study utilizes both textual and numerical data. Textual data are in the form of financial news articles and were obtained from the Financial News and Stock Price Integration Dataset (FNSPID)

Data Set. These data are publicly available at: https://github.com/Zdong104/FNSPID_Financial_News_Dataset/tree/main/data_processor/gpt_sentiment_price_news_integrate. Numerical data consisted of historical share prices of stocks in the portfolio and were obtained from Yahoo Finance for the duration of the testing period.

3.1.1. Financial text data

The FNSPID dataset consists of news detailing various events, developments, and sentiment in the market, including, but not limited to, companies' earnings- performance results, analysts' predictions and estimates, trading activity, and general economic data (e.g., retail sales data, quarterly economic growth, monetary policy stance, and expectations). While most of the articles were end-of-day, some were written during the trading day to highlight key market developments as they transpired. The dataset, which includes 29.7 million stock prices and 15.7 million financial news records for 4,775 S&P 500 companies from 1999 to 2023, was divided into training, validation, and testing sets using a 50% – 30% – 20% split. The first 50% of the data (1999 - 2010) was allocated for training, the next 30% (2011–2017) for validation, and the remaining 20% (2018–2023) was reserved for testing. This split ensures that the model is trained on a substantial portion of historical data, fine-tuned on a separate validation set, and evaluated on a completely unseen testing set to simulate real-world conditions. During the training phase, the LSTM model processes sentiment scores generated by three individual LLMs as sequential inputs, learning to capture temporal dependencies and relationships between these scores.

The LSTM architecture was configured with one layer and 50 hidden units, as shown in Table 1, which provided a balance between model complexity and computational efficiency. To mitigate over-fitting and ensure robust generalization to unseen data, a dropout rate of 0.2 and L2 weight regularization of $1e^{-4}$ were applied. The model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 50 epochs, which allowed for stable convergence while avoiding excessive training time. The Adam optimizer's default parameters were used to ensure efficient optimization. These configurations, as detailed in Table 1, were carefully chosen to enhance the model's ability to aggregate sentiment scores effectively while maintaining computational efficiency and generalization performance. Finally, the model was evaluated on the testing set (2018–2023) to assess its generalization ability and ensure unbiased performance metrics. This structured approach ensures robust model training, effective hyperparameter tuning, and reliable evaluation, ultimately enhancing the accuracy of sentiment score aggregation and its application in portfolio optimization.

http://dx.doi.org/10.21511/imfi.22(3).2025.16

Table 1. Configured parameters for LSTM sentiment aggregation

Parameter	Configured Value
Network Architecture Parameters	
Number of LSTM layers	1
Number of hidden units	50
Regularization Parameters	
Dropout rate	0.2
L2 weight regularization	$1e^{-4}$
Training Parameters	
Learning rate	0.001
Batch size	32
Number of training epochs	50
Optimizer	Adam
Adam Optimizer Parameters	
Beta 1	0.9
Beta 2	0.999
Epsilon	$1e-8$
Weight Decay	0.0
Amsgrad	False

3.1.2. *Meta-LLM sentiment scores*

Meta-LLM sentiment scores are generated by aggregating outputs from three LLMs processing FNSPID financial text data, using an LSTM network selected for its superior performance over alternative models (Random Forest, Gradient Boosting, SVMs). The LSTM was trained on the first 25% of the dataset's LLM-derived sentiment scores and validated on the remaining data, demonstrating optimal predictive accuracy through lower error metrics (MSE, MAE) and higher explanatory power (R^2) as shown in Table 2. These results are consistent with the hypothesis that states that the LSTM aggregation method produces sentiment scores with higher accuracy compared to alternative machine learning methods when integrating multi-LLM outputs ($H03_{Alternative}$).

Table 2. Machine learning models' predictive performance

Model	MSE	MAE	R^2
Random Forest	0.0304	0.4107	0.991288
Gradient Boosting	0.0309	0.6683	0.991162
Support Vector Machines	0.5055	7.0393	0.855335
LSTM	0.0005	0.1376	0.999869

Using Meta-LLM sentiment scores can be an effective strategy to mitigate divergence in individual LLM sentiment scores and reduce the impact of hallucinations. Table 3 highlights the advantage of using an aggregate or Meta-LLM sentiment score instead of sentiment scores based on individual LLMs by showcasing the performance variability among individual LLMs used for sentiment analysis. The table highlights the poor performance of individual LLMs in accurately capturing sentiment compared to the Meta-LLM model, where the accurate sentiment score or ground truth label is the average of the sentiment scores from the three LLMs adjusted for accuracy[3]. This underperformance is attributed to model-specific limitations and LLM hallucinations.

A Meta-LLM model based on LSTM aggregation mitigates these issues through combining the outputs of individual LLMs, averaging out inconsistencies, and reducing the impact of hallucinations. The use of LSTM (which is designed to capture long-term dependencies), in aggregating sentiment, provides a more contextually aware and consistent sentiment output. As a result, the Meta-LLM aggregation model achieves significantly improved performance metrics, including a 0.94 F1 Score and 0.97 accuracy. By leveraging the complementary strengths of individual models and addressing their limitations, the Meta-LLM demonstrates the practical value of ensemble techniques in producing reliable, high-quality sentiment predictions, which can be effectively used for portfolio optimization.

Table 3. LLMs' performance metrics on sentiment analysis

Sentiment Model	F1 Score	Accuracy
FinBERT ProsusAI	0.64	0.76
FinBERT Yiyanghkust	0.66	0.86
Distil Roberta	0.63	0.85
Meta-LLM	0.94	0.97

3.1.3. *Numerical data*

Although textual data proved valuable in estimating sentiment scores for incorporation into portfolio optimization, its utility is limited when employed in isolation, particularly in the context of this study, which leverages the Black-Litterman

[3] The ground truth label is derived by averaging the sentiment scores from the three individual LLMs, with an adjustment for accuracy. This adjustment is crucial to ensure that the LSTM-based Meta-LLM aggregation model not only combines the scores from the individual LLMs but also produces an aggregated value that accurately reflects the sentiment expressed in the text.

model to integrate sentiment-based views into portfolio optimization. The Black-Litterman framework, being fundamentally quantitative, requires numerical inputs, including the variance-covariance matrix and expected returns derived from historical stock price data. This computational requirement necessitates the incorporation of numerical market data, such as share prices, to effectively complement the text-derived sentiment scores within the proposed methodology.

A diversified 10-stock portfolio from the S&P 500 Index is selected based on mention frequency in the FNSPID Financial News Data Set (2018–2023) to maximize sentiment data utilization. The top-mentioned stocks are as follows: Microsoft (MSFT), Apple (AAPL), JPMorgan Chase (JPM), Oracle (ORCL), Bank of America (BAC), Amazon (AMZN), BHP Group (BHP), Cisco Systems (CSCO), PepsiCo (PEP), and Tesla (TSLA). Table 4 presents the descriptive statistics for these constituents during the testing period. Historical returns were used for performance benchmarking.

Table 4. Equity returns descriptive statistics

Ticker	Mean (%)	Std Dev (%)	Skewness	Kurtosis
AAPL	0.12	1.83	–0.01	5.79
AMZN	0.11	2.06	0.21	5.35
BAC	0.08	2.02	0.33	10.68
BHP	0.09	2.12	–0.18	5.44
CSCO	0.05	1.61	–0.58	11.91
JPM	0.08	1.75	0.34	14.97
MSFT	0.12	1.73	0.00	7.65
ORCL	0.07	1.73	0.57	20.82
PEP	0.04	1.21	–0.20	22.88
TSLA	0.21	3.65	0.16	4.24

Table 4 reveals all portfolio stocks generated positive daily mean returns (2018–2023), ranging from PepsiCo's (PEP) stable 0.05% to Tesla's (TSLA) high-growth 0.21%. Volatility varied significantly, with TSLA showing the highest standard deviation (3.65%) versus PEP's 1.21%, demonstrating the classic risk-return trade-off. The return distributions exhibited strong leptokurtosis (ORCL: 20.82, PEP: 22.88), indicating heightened tail risk during market stress. Skewness patterns diverged: BHP and CSCO showed negative skewness (frequent small gains/rare large losses), while ORCL and BAC displayed positive skewness (frequent small losses/rare large gains), offering investors distinct risk-return profiles to match their preference.

3.2. Portfolio performance

Improving the sentiment of a stock together will lead to high expected returns through a higher predicted share price, which will lead to an increase in the weight of the stock in the portfolio, while the opposite is also true. Changes in portfolio weights over time will influence the performance of the LSTM Meta-LLM Black-Litterman portfolio compared to other portfolios. The performance of the LSTM Meta-LLM Black-Litterman portfolio is first compared against benchmark portfolios, including the traditional Black-Litterman approach. Subsequently, the portfolio's performance is evaluated against alternative portfolios constructed using sentiment scores from individual LLMs rather than the aggregated Meta-LLM sentiment scores.

To evaluate portfolio performance relative to benchmark alternatives, comparisons are made between the proposed portfolio and three benchmark strategies: (1) the traditional Black-Litterman portfolio, (2) an Equally Weighted Portfolio, and (3) a Market Capitalization Weighted Portfolio (Cap Weighted Portfolio). The traditional Black-Litterman portfolio applies the same portfolio optimization process as the LSTM Meta-LLM Black-Litterman portfolio without incorporating Meta-LLM sentiment scores derived from LLMs in the optimization process. This means the Traditional Black-Litterman portfolio is fully reliant on historical share price data and does not combine subjective views of investors in the portfolio optimization.

Table 5 shows the performance of the LSTM Meta-LLM Black-Litterman portfolio compared to the benchmark portfolios over the testing period. The table highlights the great performance of the LSTM Meta-LLM Black-Litterman portfolio compared to the Traditional Black-Litterman Portfolio, Equally Weighted Portfolio, and the Cap Weighted Portfolio, with the LSTM Meta-LLM Black-Litterman portfolio achieving the highest annualized return of 31.22% and the highest risk-adjusted performance in the form of the Sharpe

http://dx.doi.org/10.21511/imfi.22(3).2025.16

Table 5. Portfolio performance comparison

Portfolio	Annualized Return (%)	Sharpe Ratio	Omega Ratio	Jensen's Alpha (%)
LSTM Meta-LLM Black-Litterman	31.22	3.02	2.48	1.95
Traditional Black-Litterman	30.11	2.80	2.31	1.87
Cap Weighted	24.57	2.37	2.26	1.56
Equally Weighted	27.28	2.58	2.24	1.71

ratio[4] of 3.02 and the highest Omega Ratio[5] of 2.48. The LSTM Meta-LLM Black-Litterman portfolio also produces positive and the highest Jensen's Alpha, indicating that not only does this portfolio outperform the market, but it also outperforms other benchmark portfolios.

These results suggest that the LSTM Meta-LLM Black-Litterman approach significantly enhances portfolio performance through its ability to integrate sentiment analysis from LLMs with traditional asset allocation techniques such as the Black-Litterman model. The high Sharpe and Omega ratios indicate that the LSTM Meta-LLM Black-Litterman portfolio not only achieves superior returns but does so with lower risk when compared to other strategies considered. Additionally, the use of Meta-LLM sentiment scores in portfolio optimization resulted in higher and positive Jensen's Alpha, thereby confirming that this method is a robust active management strategy that can capitalize on market inefficiencies. These empirical results validate the significant potential of integrating advanced machine learning (LSTM-based Meta-LLM sentiment) with traditional portfolio optimization frameworks. The findings support the alternative hypothesis ($H01_{Alternative}$) that sentiment-augmented portfolios generate superior absolute and risk-adjusted returns compared to conventional benchmarks.

Meta-LLM sentiment scores are incorporated into the portfolio optimization framework by interpreting them as investor views in the Black-Litterman portfolio optimization process. The confidence placed on these investor views is controlled by the τ parameter. As described in Equation 3, the τ parameter measures uncertainty in the prior estimate of returns, therefore also measuring uncertainty or confidence in the investor's view. By default, a τ value of 0.5 is used, which represents a moderate balance between market equilibrium and the investor's sentiment-driven expectations (based on Meta-LLM sentiment scores). An increase in τ results in greater weighting of sentiment-based views, enhancing the model's responsiveness to LLM-generated sentiments while reducing its dependence on market equilibrium expected returns. Table 6 shows that increasing the value of τ leads to higher returns and improved risk-adjusted returns. This highlights the positive impact that LLM sentiments have on generating better portfolio returns, placing more confidence in LLM sentiments produces improved portfolio performance.

Table 6. Portfolio performance comparison: Varying τ parameter

τ	Annualized Return (%)	Sharpe Ratio	Omega Ratio	Jensen's Alpha (%)
0.05	31.11	2.96	2.47	1.93
0.5	31,22	3.02	2.48	1.95
0.8	31.31	3.06	2.49	1.96

While it has been established that the LSTM-Meta-LLM Black-Litterman portfolio performs better than both optimized portfolios (Traditional Black-Litterman portfolio) and naive portfolios (Equally Weighted and Cap Weighted portfolios) using different performance measures, it is important to understand how the portfolio performs compared to the 3 individual LLMs. The portfolio performance of each individual LLM portfolio was calculated using the same methodology as the LSTM Meta-LLM Black-Litterman portfolio; however, instead of using the LSTM-based Meta-Sentiment scores, actual sentiment scores from each of the three LLMs were used to inform views in the Black-Litterman portfolio optimization. Table 7 shows how the LSTM-Meta-LLM Black-Litterman performs against the individual LLM models.

[4] Sharpe *Ratio* = μ/σ where μ and σ are the mean and standard deviation of the portfolio returns.

[5] Omega ratio = μ^u/μ^d, where μ^u is the absolute deviations above the benchmark and μ^d is the absolute deviations below the benchmark (A performance benchmark return of 0% is assumed to calculate Omega).

Table 7. Portfolio performance: Meta-LLM vs individual LLMs

LLM Portfolio	Annualized Return (%)	Sharpe Ratio	Omega Ratio	Jensen's Alpha (%)
LSTM Meta-LLM Black-Litterman	31.22	3.02	2.48	1.95
FinBERT ProsusAI	34.68	2.59	2.13	2.26
FinBERT Yiyanghkust	30.37	2.43	2.21	1.91
Distil Roberta	33.48	2.63	2.07	2.00
Meta-LLM Rank	3rd	1st	1st	3rd

The Meta-LLM Model demonstrates superior risk-adjusted performance compared to individual LLMs. Notably, it has achieved the highest Sharpe Ratio and the Omega ratio, while its Jensen's Alpha ranks third. This outstanding performance is a testament to the model's ability to effectively combine the strengths of individual LLMs, resulting in more robust and accurate sentiment scores. By leveraging the unique advantages of each model, the Meta-LLM Model can produce a more comprehensive and reliable sentiment analysis, ultimately leading to better risk-adjusted portfolio performance.

While the LSTM Meta-LLM Black-Litterman portfolio's annualized return of 31.22% modestly trails the 34.68% achieved by the top-performing individual LLM (FinBERT ProsusAI), this differential reflects the model's deliberate risk-management strategy. The FinBERT approach generates higher returns at the cost of significantly greater volatility (Sharpe Ratio = 2.59 vs. the LSTM Meta-LLM Black-Litterman portfolio's 3.02), confirming that the LSTM-based aggregation successfully reduces sentiment volatility through cross-model bias mitigation, a finding that strongly supports the alternative hypothesis, $H02_{Alternative}$. This performance profile demonstrates that the Meta-LLM framework transforms the unstable outputs of individual LLMs into consistent excess returns, achieving the optimal balance between performance and stability for practical portfolio management. Overall, the results suggest that the Meta-LLM Black-Litterman Model is a valuable tool for investors seeking to optimize their portfolios and make more informed investment decisions.

CONCLUSIONS

This study has successfully developed and validated an advanced portfolio optimization framework that integrates machine learning-enhanced sentiment analysis with traditional asset allocation methods. By combining sentiment scores from three specialized LLMs through an LSTM network and incorporating them as dynamic views in the Black-Litterman model, the study demonstrated significant improvements in both portfolio returns and risk management. The key empirical finding shows that the LSTM Meta-LLM Black-Litterman portfolio achieved a 31.22% annualized return with a Sharpe Ratio of 3.02, outperforming traditional benchmarks in both absolute returns and risk-adjusted performance.

Three critical insights emerge from these results. First, the multi-LLM approach proves essential for reliable sentiment analysis, as it reduces single-model volatility while capturing complementary perspectives from different training corpora. Second, the τ parameter serves as a powerful calibration tool, allowing investors to systematically balance sentiment-driven views with market equilibrium expectations, empirical testing demonstrates that increasing τ consistently improves performance. Third, the framework's success in generating positive Jensen's Alpha (1.95%) confirms its active management potential, capable of identifying sentiment-driven market inefficiencies that traditional models miss.

These findings have transformative implications for investment practice. Asset managers can implement this approach to enhance existing quantitative strategies with AI-driven sentiment signals while maintaining the risk control of classical portfolio theory. The methodology is particularly valuable for: (1) actively managed funds seeking consistent alpha, (2) risk-averse investors requiring stable returns, and (3) quantitative teams looking to modernize traditional factor models. Future extensions could explore

http://dx.doi.org/10.21511/imfi.22(3).2025.16

real-time implementation with high-frequency sentiment data or applications to multi-asset portfolios, though careful consideration of transaction costs remains essential. Ultimately, this study establishes a new paradigm where NLP and portfolio theory interact synergistically, machine learning extracts nuanced signals from financial texts, while time-tested optimization methods ensure their prudent implementation in investment portfolios.

AUTHOR CONTRIBUTIONS

Conceptualization: Lamukanyani Alson Mantshimuli, John Weirstrass Muteba Mwamba.
Data curation: Lamukanyani Alson Mantshimuli.
Formal analysis: Lamukanyani Alson Mantshimuli.
Investigation: Lamukanyani Alson Mantshimuli, John Weirstrass Muteba Mwamba.
Methodology: Lamukanyani Alson Mantshimuli, John Weirstrass Muteba Mwamba.
Project administration: Lamukanyani Alson Mantshimuli.
Resources: Lamukanyani Alson Mantshimuli.
Software: Lamukanyani Alson Mantshimuli.
Supervision: John Weirstrass Muteba Mwamba.
Validation: Lamukanyani Alson Mantshimuli.
Visualization: Lamukanyani Alson Mantshimuli.
Writing – original draft: Lamukanyani Alson Mantshimuli.
Writing – review & editing: John Weirstrass Muteba Mwamba.

REFERENCES

1. Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models.* Retrieved from https://arxiv.org/abs/1908.10063
2. Benjamin, J., & Mathew, J. (2025). Enhancing continuous integration predictions: a hybrid LSTM-GRU deep learning framework with evolved DBSO algorithm. *Computing, 107*(1), 9. Retrieved from https://link.springer.com/article/10.1007/s00607-024-01370-2
3. Black, F., & Litterman, R. (1990). Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income, 1,* 7-18. Retrieved from https://www.scirp.org/reference/referencespapers?referenceid=2912782
4. Bukhari, A.H., Raja, M.A.Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access, 8,* 71326-71338. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9057460
5. Colasanto, F., Grilli, L., Santoro, D., & Villani, G. (2022). BERT's sentiment score for portfolio optimisation: a fine-tuned view in Black and Litterman model. *Neural Computing and Applications, 34*(20), 17507-17521. Retrieved from https://link.springer.com/article/10.1007/s00521-022-07403-1
6. de Kok, S., Punt, L., van den Puttelaar, R., Ranta, K., Schouten, K., & Frasincar, F. (2018). Review-Aggregated aspect-based sentiment analysis with ontology features. *Progress in Artificial Intelligence, 7,* 295-306. Retrieved from https://link.springer.com/article/10.1007/s13748-018-0163-7
7. Dmonte, A., Ko, E., & Zampieri, M. (2024, December). An Evaluation of Large Language Models in Financial Sentiment Analysis. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 4869-4874). IEEE. Retrieved from https://ieeexplore.ieee.org/document/10825272
8. Dong, Z., Fan, X., & Peng, Z. (2024, August). Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 4918-4927). Retrieved from https://arxiv.org/abs/2402.06698
9. Huang, D., Huang, K., Li, Z., Liu, Z., & Zhao, J. (2021, January). FinBERT: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 45134519). Retrieved from https://dl.acm.org/doi/abs/10.5555/3491440.3492062
10. Kang, H., & Liu, X. Y. (2023). *Deficiency of large language models in finance: An empirical examination of hallucination.* Retrieved from https://ideas.repec.org/p/arx/papers/2311.15548.html
11. Kirtac, K., & Germano, G. (2025). Leveraging LLM-based sentiment analysis for portfolio allocation with proximal policy optimization. In *ICLR 2025 Workshop on Machine Learning Multiscale Processes.* Retrieved from https://aclanthology.org/2025.realm-1.12/
12. Kuruvilla, J. S., & Mythily, M. (2025). Financial LLM For Stock

Price Analysis And Investment Recommendation. *Journal of Telematics and Informatics, 13*(1). Retrieved from https://section.iaesonline.com/index.php/JTI/article/view/6316

13. Lefort, B., Benhamou, E., Ohana, J. J., Saltiel, D., & Guez, B. (2024). *Optimizing Performance: How Compact Models Match or Exceed GPT's Classification Capabilities through Fine-Tuning.* Retrieved from https://ideas.repec.org/p/arx/papers/2409.11408.html

14. Liu, B. (2012). *Sentiment analysis and opinion mining.* Springer Nature. Retrieved from https://link.springer.com/book/10.1007/978-3-031-02145-9

15. Lou, R., Zhang, K., & Yin, W. (2024). Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics, 50*(3), 1053-1095. https://doi.org/10.1162/coli_a_00523

16. Mao, D., Zhang, D., Zhang, A., & Zhao, Z. (2025, April). MLS-DET: Multi-LLM Statistical Deep Ensemble for Chinese AI-Generated Text Detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. Retrieved from https://ieeexplore.ieee.org/document/10888686

17. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. Retrieved from https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

18. Pathuri, S. K., Anbazhagan, N., & Prakash, G. B. (2020, July). Feature based sentimental analysis for prediction of mobile reviews using hybrid bag-boost algorithm. In *2020 7th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-5). IEEE. Retrieved from https://ieeexplore.ieee.org/document/9201990

19. Ranjan, R., Gupta, S., & Singh, S. N. (2024). *A comprehensive survey of bias in LLMs: Current landscape and future directions.* Retrieved from https://www.researchgate.net/publication/384363990_A_Comprehensive_Survey_of_Bias_in_LLMs_Current_Landscape_and_Future_Directions

20. Romero, M. (2024). *Distil Roberta.* Retrieved from https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

21. Shah, N., Genc, Z., & Araci, D. (2024). *StackEval: Benchmarking LLMs in Coding.* https://doi.org/10.48550/arXiv.2412.05288

22. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In Chinese computational linguistics. *18th China National Conference, CCL 2019.* Kunming, China (pp. 194-206). Springer International Publishing. Retrieved from https://arxiv.org/abs/1905.05583

23. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & Rodriguez, A. (2023). *LLaMA: Open and efficient foundation language models.* Retrieved from https://www.researchgate.net/publication/368842729_LLaMA_Open_and_Efficient_Foundation_Language_Models

24. Vaswani, A. (2017). *Attention is all you need. Advances in Neural Information Processing Systems.* Retrieved from https://arxiv.org/abs/1706.03762

25. Wang, J. H., Liu, T.W., Luo, X., & Wang, L. (2018, October). An LSTM approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)* (pp. 214-223). Retrieved from https://aclanthology.org/O18-1021.pdf

26. Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., & Xu, Y. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems, 37,* 95716-95743. https://dl.acm.org/doi/10.5555/3737916.3740949

27. Zhang, B., Yang, H., & Liu, X.Y. (2023). *Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models.* Retrieved from https://ideas.repec.org/p/arx/papers/2306.12659.html

28. Zhang, H., & Shafiq, M.O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data, 11*(1), 25. Retrieved from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00842-0

29. Zhang, W.E., Sheng, Q.Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST), 11*(3), 1-41. https://doi.org/10.48550/arXiv.1901.06796

30. Zhao, H., Liu, Z., Wu, Z., Li, Y., Y., T., Shu, P., Xu, S., Dai, H., Zhao, L., Mai, G., & Liu, N. (2024). *Revolutionizing finance with LLMs: An overview of applications and insights.* https://doi.org/10.48550/arXiv.2401.11641