Benchmarking Large Language Models for Target-Based Financial Sentiment Analysis

Iftikhar Muhammad[1,*], Marco Rospocher[1], Timotej Knez[2] and Slavko Žitnik[2]

[1]University of Verona, 37129 Verona, Italy
[2]University of Ljubljana, 1000 Ljubljana, Slovenia

Abstract

Sentiment analysis is vital for understanding market dynamics and formulating informed investing strategies, especially in volatile financial conditions. This study advances target-based financial sentiment analysis (TBFSA) by rigorously evaluating the efficacy of Large Language Models (LLMs) in zero-shot and few-shot learning contexts. We compare cutting-edge generative LLMs, such as ChatGPT-4o, ChatGPT-4, ChatGPT-o1, DeepSeek-R1, Llama-3-8B, Gemma-2-9B, and Gemma-2-27B, with conventional lexicon-based tools (VADER, TextBlob) and discriminative transformer-based models (FinBERT, FinBERT-Tone, DistilFinRoBERTa, Deberta-v3-base-absa-v1.1). Our analysis utilizes a newly curated dataset of 1,162 manually annotated Bloomberg news articles, designed explicitly for TBFSA (due to copyright constraints, only URLs are publicly released, with full news content accessible through a Bloomberg Terminal). The findings indicate that LLMs, particularly DeepSeek-R1 and ChatGPT variants (especially ChatGPT-o1), outperform lexicon-based approaches and discriminative transformer-based models across all evaluation metrics, without requiring additional training or task-specific fine-tuning. The study establishes generative LLMs as a scalable and cost-effective method for target-level sentiment analysis, relieving the need for expensive, rigorous fine-tuning. The research provides valuable insights, enabling institutions to use unstructured textual data effectively for improved real-time risk assessment, portfolio management, and algorithmic trading.

Keywords

Large Language Models, Target-Based Sentiment Analysis, Financial Sector

1. Introduction

The financial sector, a pivotal pillar of the global economy, is increasingly influenced by vast amounts of unstructured textual data, including news articles, earnings call transcripts, regulatory filings, and analyst reports [1]. These textual sources significantly impact investor decisions, market volatility, and strategic financial activities [2]. The inadequacy of traditional manual methods for processing such extensive data has led to adopting automated procedures using Natural Language Processing (NLP) techniques [3]. Sentiment analysis, a crucial NLP tool, evaluates the emotional tone of the text, providing valuable predictive insights on investor sentiment and market movements [2].

Financial Sentiment Analysis (FSA), a specific subtask of NLP, identifies subjective tones in financial texts, offering insights for market forecasting, risk management, and the development of trading strategies [4]. Methods for FSA range from conventional lexicon-based techniques and machine learning algorithms to advanced deep learning models, particularly transformer architectures [5]. Recently, generative large language models (LLMs) such as Llama, Gemma, ChatGPT, and DeepSeek have exhibited considerable promise in NLP tasks, especially in zero-shot and few-shot learning contexts, owing to their ability to reduce reliance on extensive manual annotations [6]. However, the efficacy of these models in specialized fields, such as finance, is still inadequately examined, underscoring the necessity for thorough assessment before their incorporation into practical applications like financial reporting software and trading algorithms.

A notably complex facet of sentiment analysis in financial texts is the recurrent presence of conflicting sentiments towards multiple entities within a single narrative [7]. For example, the statement "Nvidia's AI-driven growth overshadows Netflix's subscriber stagnation" concurrently expresses positive and negative sentiments regarding two distinct entities. Conventional sentiment analysis methods at the sentence or document level frequently conflate these subtle perspectives, obscuring critical insights necessary for precise decision-making. To overcome this constraint, Target-Based Financial Sentiment Analysis (TBFSA) disaggregates sentiment at the entity level, facilitating a more detailed examination of specific financial instruments, business entities, or market segments [8]. Nonetheless, the capacity of LLMs to execute zero-shot and few-shot TBFSA tasks in finan-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
iftikhar.muhammad@univr.it (I. Muhammad); marco.rospocher@univr.it (M. Rospocher); timotej.knez@fri.uni-lj.si (T. Knez); slavko.zitnik@fri.uni-lj.si (S. Žitnik)
0000-0003-4747-9680 (I. Muhammad); 0000-0001-9391-3201 (M. Rospocher); 0000-0001-7506-5739 (T. Knez); 0000-0003-3452-1106 (S. Žitnik)

cial markets remains insufficiently investigated. Furthermore, rigorous comparison analyses of lexicon-based tools, discriminative transformer-based approaches, and generative LLMs in this particular setting remain scarce.

The current study aims to fill these significant gaps by evaluating the potential of LLMs to conduct target-specific sentiment analysis in financial news articles. Specifically, we seek to answer the following research questions:

1. How do zero-shot and few-shot generative LLMs perform in TBFSA compared to lexicon-based and discriminative transformer-based models?

2. Does few-shot learning substantially improve the performance of LLMs compared to zero-shot methods in TBFSA?

Our contributions can be summarized as follows:

1. We develop and publicly release a novel, manually annotated TBFSA dataset comprising 1,162 financial news articles categorized by target-specific sentiments. In contrast to current financial datasets (e.g., FiQA-2018,[1] Financial PhraseBank [9]), our dataset distinctly encapsulates sophisticated entity-level opinions within intricate financial narratives that exhibit conflicting sentiments.

2. Utilizing this dataset, we systematically evaluate generative LLMs (ChatGPT, Llama, Gemma, DeepSeek), conventional lexicon-based instruments (VADER, TextBlob), and discriminative transformer-based models (Finbert, DistilFinRoBERTa, Finbert-Tone, Deberta-v3-base-absa-v1.1), emphasizing the strengths and limitations of each approach specifically in the context of TBFSA. This extensive comparison investigation is among the first to critically evaluate advanced LLMs' performance in zero-shot and few-shot frameworks for target-level financial sentiment analysis.

The subsequent sections of this research are structured as follows: Section 2 presents relevant literature on financial sentiment analysis. Section 3 delineates the establishment of our dataset, annotation processes, and methodological techniques. Section 4 delineates empirical findings and discussion, while Section 5 concludes the study and provides key implications and avenues for future research.

[1] https://sites.google.com/view/fiqa/home

2. Related Work

2.1. Lexicon-Based Methods

Lexicon-based approaches, which form the foundation of financial sentiment analysis, initially drew from general-purpose instruments such as LIWC and SentiWordNet. However, these tools lacked domain-specific accuracy and contextual nuance [10]. Frameworks like VADER and TextBlob were then developed to incorporate contextual scoring and automatic lexicon enhancement [11, 12]. Numerous scholars have utilized VADER in the financial domain [13, 14, 15]. However, it struggles to handle sector-specific terminology [16]. Similarly, TextBlob, which integrates predefined lexicons with a classifier trained on film reviews, allows for swift implementation in initial analyses. However, it falls short in complex financial scenarios due to its inadequate domain adaptation [16].

While lexicon-based methods have been practical, they face significant challenges in deciphering complex linguistic patterns, domain-specific vocabulary, and contextual nuances [17]. These limitations have led to transformer-based models leveraging deep learning to capture semantic and contextual subtleties more effectively in financial texts.

2.2. Discriminative Transformer-Based Models

Transformer-based architectures, particularly BERT [18], transformed NLP by employing a self-attention technique that effectively captures contextual relationships. Although general transformers excel at conventional NLP tasks, their effectiveness declines in financial contexts due to specialized lexicons and nuanced tone differences. As a result, domain-specific models fine-tuned on financial data have developed an increased sensitivity to the subtleties of financial language and numerical settings [19].

FinBERT [20], trained initially on financial documents like SEC filings and subsequently fine-tuned with the FiQA dataset, represented a notable progression in financial sentiment analysis. Studies conducted by [19, 21] confirmed FinBERT's superiority compared to general-purpose models, especially in analyzing earnings transcripts. Expanding on this, FinBERT-Tone [22] implemented tonal analysis to discern subtle sentiment indications essential for market forecasting. Initiatives to improve efficiency, shown by DistilFinRoBERTa [23], tailored for real-time applications, have also garnered attention. Furthermore, sophisticated models like DeBERTa-v3-base-absa-v1.1 exhibited accuracy in aspect- and target-oriented sentiment analysis, adeptly interpreting intricate narratives in financial documents [17].

Comparative assessments consistently demonstrate that fine-tuned transformer-based models exceed traditional lexicon-based and machine-learning methodologies [19, 24]. Nevertheless, their demand for processing resources and extensive labelled datasets has initiated the exploration of generative LLMs as viable alternatives that scale more effectively with fewer task-specific labels.

2.3. Generative Large Language Models

Recent developments in LLMs have shown exceptional proficiency in FSA, surpassing conventional lexicon-based and discriminative transformer-based methodologies [21]. The intricate linguistic characteristics of financial texts have prompted the creation of specialized LLMs, such as BloombergGPT [25] and FinVis-GPT [26], specifically tailored for the financial sector. Models such as InvestLM [27], especially fine-tuned for investing environments, have demonstrated effectiveness equivalent to commercial advice systems.

Furthermore, recent research highlights the efficacy of smaller, computationally efficient models, attaining performance akin to larger LLMs via focused fine-tuning. Methods like parameter-efficient tuning (e.g., LoRA) have enhanced their utilization in practical financial scenarios [28]. Significantly, even general-purpose models such as ChatGPT have exhibited remarkable proficiency in financial sentiment analysis without the necessity for domain-specific fine-tuning [29].

Despite significant progress, previous studies have primarily focused on generic sentiment analysis, with limited investigation into target-based sentiment analysis within financial contexts. While [17] examined the zero-shot efficacy of LLMs on financial headlines, our research expands this investigation by evaluating full-text articles to provide more extensive contextual insights. Additionally, we extend the evaluation framework to encompass few-shot scenarios and a varied array of models—such as Llama 3-8B, Gemma 2 (9B and 27B), DeepSeek-R1, and ChatGPT variants—benchmarked against conventional lexicon-based and discriminative transformer-based models. Unlike [17], which examined sentiment toward a single target per headline, our study investigates multiple targets within each article, enabling more granular and comprehensive financial sentiment analysis.

3. Methodology

This section delineates the methodological framework utilized to assess the performance of generative LLMs—specifically, Gemma, Llama, ChatGPT, and DeepSeek—in executing TBFSA. To effectively benchmark these LLMs, we utilized various lexicon-based sentiment analysis tools, specifically VADER and TextBlob, in conjunction with discriminative transformer-based models, including FinBERT, DistilFinRoBERTa, FinBERT-Tone, and DeBERTa-v3-base-absa-v1.1. We began by outlining our methodology for dataset collecting and annotation, a meticulous process that ensured high reliability and validity criteria. Subsequently, we fine-tuned the benchmark discriminative transformer-based models utilizing this dataset to achieve optimal alignment with the specific requirements of financial sentiment analysis. To thoroughly assess the generative LLMs, we designed precise, task-oriented prompts appropriate for TBFSA. Finally, we conducted a comprehensive comparative study to evaluate the efficacy and robustness of LLMs compared to the benchmark models.

3.1. Dataset Construction and Annotation

To establish a thorough evaluation framework, we obtained news articles from the Bloomberg Terminal regarding four prominent stock companies—Alphabet, Amazon, Netflix, and Nvidia. The assembled dataset comprises 1,170 articles dated from September 4, 2023, to January 30, 2024. Each article was systematically analyzed to extract critical information, including the timestamps, news text (excluding headlines), and URLs, which were then organized in a structured database (as depicted in Figure 1).

Each article was meticulously annotated for sentiment concerning the target companies to ensure data quality and confirm the experimental evaluation. The annotation was carried out by three annotators with extensive expertise in finance and economics, all possessing advanced English competence (CEFR level C1). Their annotations were guided by comprehensive guidelines aimed at standardizing target identification and sentiment assessment.

A concise summary of these guidelines entails:

- A thorough examination of each article to identify direct references to the target entities: Alphabet, Amazon, Netflix, and Nvidia.
- Identification of multiple target entities within a single article, where applicable.
- Labelling articles devoid of explicit target references as “no target.”
- Evaluation of sentiment from an investor’s viewpoint, relying exclusively on the textual content.
- Sentiment classification as positive (1), negative (-1), or neutral (0).
- Identification of prevailing sentiment in instances of mixed expressions.
- Neutral labelling for vague, ambiguous, or passing references.

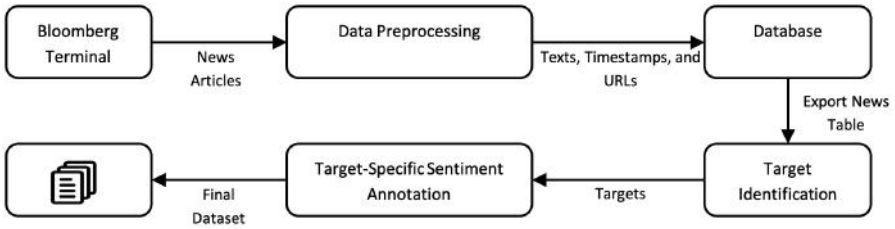

Figure 1: Dataset construction process.

The annotating procedure was organized into two separate phases. The annotators initially conducted target identification individually across all 1,170 articles. Eight articles were excluded as having "no target" by consensus. Inter-annotator reliability for target identification yielded a Krippendorff's alpha [30] of 0.96 and a percentage agreement [31] of 98.95% for the remaining 1,162 articles, signifying consistent annotations. Texts with majority-agreed targets were forwarded for sentiment annotation, yielding 1,334 unique annotation cases due to multiple target references within specific articles.

In the second phase, sentiment annotation was performed for all identified target entities. Annotators used a defined scale to assign sentiments: '1' for positive, '-1' for negative, and '0' for neutral sentiment. To ensure consistency, annotators collaboratively annotated a shared subset of 150 texts, resulting in satisfactory inter-annotator reliability (Krippendorff's Alpha of 0.81; percentage agreement of 83%). The sentiment labels for the 150 texts were established by majority consensus, and the remaining 1,184 texts were allocated evenly among annotators for individual sentiment labelling.

The final annotated dataset consists of 1,334 texts; each explicitly associated with a target entity and an annotated sentiment label. The dataset demonstrates a moderate class imbalance, with positive sentiments accounting for 45%, negative sentiments for 27%, and neutral sentiments for 28%. Table 1 presents annotated instances, whereas Figure 2 represents the sentiment distribution. Additional quantitative parameters, including the total number of news texts, average daily texts, average text length (measure in tokens), and average target mentions, are outlined in Table 2.

We publicly release our curated dataset[2] to assist the academic community and guarantee methodological transparency and reproducibility. Due to copyright restrictions, we cannot disseminate the complete content of the news articles. However, we provide comprehensive metadata, encompassing publication dates, timestamps, specified target entities, and Bloomberg article URLs, facilitating the retrieval of original articles via the Bloomberg Terminal, a subscription-based platform widely accessible in academic and financial institutions.

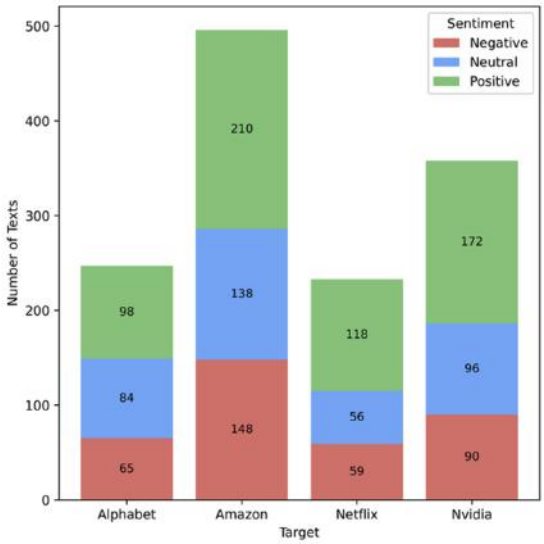

Figure 2: Sentiment distribution across the targets.

3.2. Baseline Models

To meticulously assess generative LLMs in TBFSA, we have conducted a comprehensive comparison of their efficacy with established benchmarks: lexicon-based instruments (TextBlob, VADER) and discriminative transformer architectures (FinBERT, FinBERT-Tone, DistilFinRoBERTa, DeBERTa-v3-absa-v1.1).

TextBlob,[3] an open-source python library developed on the Natural Language Toolkit (NLTK) and Pattern libraries, assigns sentiment polarity scores ranging from -1 to $+1$ and has been widely utilized for financial texts [16, 32, 33]. VADER,[4] developed by [11], a rule-based framework, incorporates lexical, grammatical, and syntactic heuristics—validated against LIWC and

[2]https://github.com/iftikharm895/Target-Based_Sentiment_Analysis_in_Financial_News

[3]https://textblob.readthedocs.io/en/dev/

[4]https://github.com/cjhutto/vaderSentiment

Table 1
Instances of annotated texts.

Target	Text	Label
Alphabet	Alphabet Inc. shares tumbled the most in a year on Wednesday after the Google parent reported a smaller than expected profit in cloud computing, raising concerns about its position in a market critical to its future. Ed Ludlow reports.	-1
Amazon	Amazon Japan says it will build its first "sort center" in Japan in Shinagawa, Tokyo, located ∼3.5km from Haneda International Airport. Expects to create ∼1,000 new jobs. Will handle as many as 750,000 items/day.	1
Netflix	Netflix co-CEO Ted Sarandos says talks with striking actors broke down after the union asked for a "levy" on streaming customers. Sarandos speaks at the first-ever Bloomberg Screentime conference in Los Angeles.	-1
Nvidia	The projected ex-date for Nvidia's dividend moved to Dec. 6 from Nov. 30, according to an updated Bloomberg Dividend Forecast. The new ex-date falls after the Dec. 1 option expiry.	0

Table 2
Dataset Statistics (the values in parentheses denote standard deviations)

Target	No of Texts	Daily Texts	Text Tokens	Target Mentions
Alphabet	247	2.84 (2.59)	456.85 (574.30)	3.29 (4.43)
Amazon	496	4.82 (3.13)	538.31 (590.15)	5.19 (6.48)
Netflix	233	3.11 (3.31)	245.92 (259.06)	3.18 (3.13)
Nvidia	358	3.81 (3.47)	381.97 (427.39)	3.32 (3.47)
Total	1334	14.58 (12.50)	430.20 (511.70)	3.99 (5.00)

ANEW—and has also been extensively employed in financial contexts [17, 34, 35].

Discriminative transformer-based baselines comprise:

1. DistilFinRoBERTa,[5] a distilled variant of RoBERTa fine-tuned on financial datasets for three-class sentiment analysis [23];
2. FinBERT[6] [20], a BERT adaptation pre-trained on earnings calls news articles and regulatory filings and fine-tuned on Financial PhraseBank [9];
3. FinBERT-Tone[7] [19], which enhances FinBERT to identify tonal nuances, fine-tuned on SEC filings, earning reports, and financial news; and
4. DeBERTa-v3-absa-v1.1,[8] builds upon the DeBERTa-v3 architecture [36], has been fine-tuned for Aspect-Based Sentiment Analysis (ABSA) through the FAST-LCF-BERT framework [37]. It is trained on an extensive dataset, comprising 30,000 ABSA-specific samples and further fine-tuned on an additional 180,000 annotated examples from a variety of datasets.

These discriminative transformer-based models have been extensively employed in financial sentiment research [23, 38, 39].

The current study involved fine-tuning DistilFinRoBERTa, FinBERT, and FinBERT-Tone using a learning rate of 3×10^{-5}, 10 training epochs, and a batch size of 32. For DeBERTa-v3-absa-v1.1, we utilized a 5-fold cross-validation approach to enhance robustness, training each fold for 10 epochs using default hyperparameters on an NVIDIA RTX 4090 GPU.

3.3. Evaluated Generative LLMs

Recent improvements in LLMs have garnered significant academic interest owing to their proven effectiveness in several text-based tasks [40]. Notable and widely utilized models include OpenAI's ChatGPT,[9] Gemma[10]—a series of open models based on Google's Gemini architecture, Meta's LLaMA,[11] and DeepSeek-[12]

The current study assessed the efficacy of various advanced generative LLMs within the framework of TBFSA. The evaluated models include ChatGPT-4, ChatGPT-4o,

[5] https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis
[6] https://huggingface.co/ProsusAI/finbert
[7] https://huggingface.co/yiyanghkust/finbert-tone
[8] https://huggingface.co/yangheng/deberta-v3-base-absa-v1.1

[9] https://chatgpt.com/. The ChatGPT variants analyzed in this study are limited to those available during the research period. Newer versions released during manuscript preparation will be examined in future work.
[10] https://gemini.google.com/app
[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[12] https://www.deepseek.com/

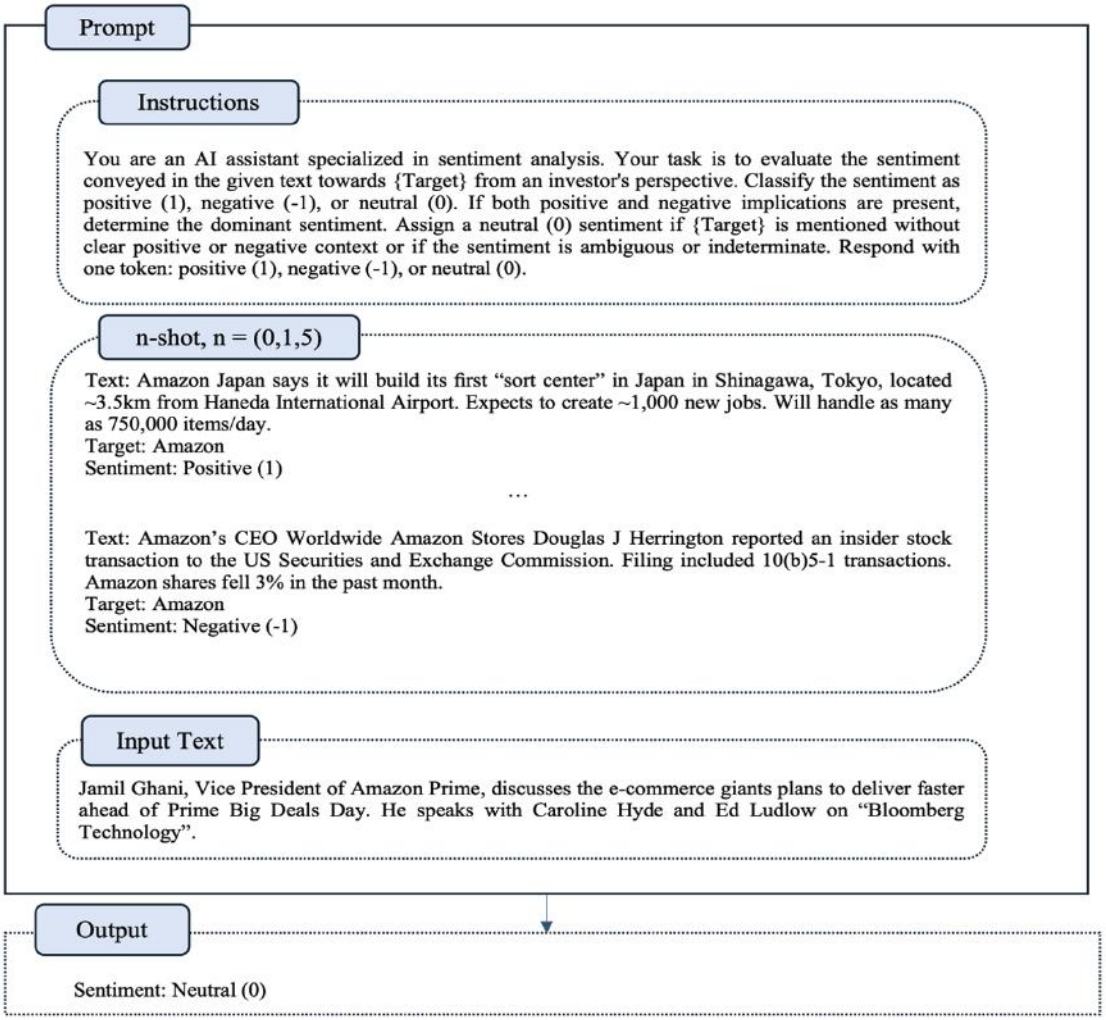

Figure 3: Prompt Used in Zero/Few-Shot Learning Approach for LLMs.

ChatGPT-o1, LLaMA 3 8B, Gemma 2 9B, Gemma 2 27B, and DeepSeek-R1. All models were assessed in their default configurations, without any additional fine-tuning, to evaluate their zero-shot and few-shot capabilities in executing the specified task. Interactions with ChatGPT variations were executed via OpenAI's standard web interface, utilizing a temperature setting of 0.7. The Gemma models were accessed via the Gemini API, which suggests a temperature setting of 1.0 for both the 9B and 27B variants. DeepSeek-R1 was accessed via its public chat interface, employing its standard temperature setting. To interact with the LLaMA model, we utilized a local instance of the Meta-Llama-3-8B-Instruct model, running under the Ollama[13] application. For testing purposes, we used default hyperparameters and advanced optimization techniques, including 4-bit quantization, to efficiently execute this model on consumer-grade GPU systems.

To assess the performance of generative LLMs in TBFSA, we employed zero-shot and few-shot prompting strategies using manually designed, fixed prompts without task-specific tuning. The prompt used in the zero/few-shot learning approach is presented in Figure 3. In the zero-shot context, models were given task instructions without illustrative examples. In few-shot contexts, prompts were augmented by either one (1-shot) or five (5-shot) additionally annotated examples, to provide contextual grounding.

This approach utilizes LLMs' inherent language and

[13]https://ollama.com/library/llama3

Table 3
Performance Outcomes of Target-Based Sentiment Classification Across Models

Model	Shot	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Weighted Precision	Weighted Recall	Weighted F1-Score
TextBlob	-	0.46	0.40	0.39	0.35	0.41	0.46	0.39
VADER	-	0.50	0.48	0.41	0.37	0.48	0.50	0.41
FinBERT	-	0.56	0.53	0.54	0.54	0.58	0.56	0.57
DistilFinRoBERTa	-	0.61	0.56	0.57	0.57	0.61	0.61	0.61
FinBERT-Tone	-	0.63	0.61	0.62	0.62	0.66	0.63	0.63
Deberta-v3-absa-v1.1	-	0.68	0.67	0.67	0.66	0.68	0.68	0.68
Llama 3 8B	0	0.68	0.75	0.62	0.63	0.72	0.68	0.66
Gemma 2 9B	0	0.66	0.69	0.65	0.66	0.71	0.66	0.67
Gemma 2 27B	0	0.69	0.70	0.68	0.69	0.71	0.69	0.70
ChatGPT-4	0	0.79	0.78	0.77	0.77	0.79	0.79	0.79
ChatGPT-4o	0	0.78	0.78	0.77	0.77	0.79	0.78	0.78
ChatGPT-o1	0	0.81	0.81	0.80	0.80	0.82	0.81	0.81
DeepSeek-R1	**0**	**0.82**	**0.84**	**0.81**	**0.81**	**0.83**	**0.83**	**0.82**
Llama 3 8B	1	0.52	0.60	0.44	0.44	0.54	0.40	0.43
Gemma 2 9B	1	0.66	0.70	0.66	0.66	0.72	0.66	0.67
Gemma 2 27B	1	0.71	0.72	0.71	0.71	0.73	0.71	0.72
ChatGPT-4	1	0.80	0.79	0.79	0.79	0.81	0.80	0.80
ChatGPT-4o	1	0.81	0.81	0.80	0.80	0.82	0.81	0.81
ChatGPT-o1	**1**	**0.85**	**0.85**	**0.84**	**0.84**	**0.85**	**0.85**	**0.85**
DeepSeek-R1	**1**	**0.83**	**0.84**	**0.82**	**0.83**	**0.83**	**0.83**	**0.83**
Llama 3 8B	5	0.64	0.69	0.61	0.63	0.64	0.64	0.63
Gemma 2 9B	5	0.65	0.68	0.65	0.65	0.71	0.65	0.66
Gemma 2 27B	5	0.72	0.71	0.71	0.71	0.73	0.72	0.72
ChatGPT-4	5	0.82	0.81	0.80	0.80	0.82	0.82	0.82
ChatGPT-4o	5	0.82	0.82	0.81	0.82	0.83	0.82	0.83
ChatGPT-o1	5	0.86	0.86	0.85	0.86	0.86	0.86	0.86
DeepSeek-R1	**5**	**0.87**	**0.88**	**0.86**	**0.87**	**0.87**	**0.87**	**0.87**

contextual reasoning abilities, enabling performance evaluation without requiring task-specific training or model adaptation. The method offers a clear assessment of model generality and adaptability, enhancing their suitability for effortless implementation in diverse practical applications.

The evaluation of model performance utilized recognized criteria for sentiment categorization, including precision, accuracy, recall, and F1-score [41]. The metrics were calculated across three sentiment categories—negative, neutral, and positive—utilizing both macro-averaging (equal weight across classes) and weighted averaging (weighted by class sample size) to ensure robustness amid moderately imbalanced class distributions, as done in analogous situations (e.g., [42]).

4. Results and Discussions

Table 3 presents the outcomes for all models evaluated on the novel dataset introduced in this research. Lexicon-based approaches, such as VADER and TextBlob, exhibit consistently subpar performance across all evaluation metrics, with macro-F1 scores below 0.37. These models are limited by their dependence on static, general-purpose sentiment lexicons that do not incorporate domain-specific financial language, in addition to their document-level emphasis and rigid rule-based architecture. As a result, they fail to capture the contextual intricacies and entity-specific sentiment differentiations necessary for effective TBFSA.

Conversely, discriminative transformer-based models optimized for FSA tasks substantially exceed the performance of lexicon-based models. FinBERT, DistilFinRoBERTa, and FinBERT-Tone attain increasingly higher macro-F1 scores (ranging from 0.54 to 0.62), demonstrating the advantages of domain-specific pretraining and contextualized embeddings. Nonetheless, these models operate at the sentence or document level and fail to assign sentiment to specific entities, hence constraining their efficacy in multi-entity financial texts. Conversely, DeBERTa-v3-base-ABSA-v1.1, tailored for target/aspect-based sentiment analysis, attains the highest macro-F1 score (0.66) among fine-tuned transformer models. Its disentangled attention mechanism and structured input encoding provide fine-grained, token-level sentiment attribution, rendering it more suitable for intricate, entity-aware financial analysis.

Among the generative LLMs evaluated under zero-shot settings, DeepSeek-R1 and the ChatGPT models (ChatGPT-o1, ChatGPT-4, and ChatGPT-4o) consistently surpass baseline models. DeepSeek-R1 attains the highest zero-shot macro-F1 score (0.82), closely followed by ChatGPT-o1 (0.80). Performance enhances with few-shot prompting: in the 1-shot setting, ChatGPT-o1 slightly outperforms DeepSeek-R1 with a macro-F1 score of 0.84 compared to 0.83. The highest scores are recorded in the 5-shot setting, with DeepSeek-R1 achieving 0.87, slightly above ChatGPT-o1's score of 0.86. These findings highlight the efficacy of few-shot learning in improving contextual comprehension and sentiment categorization outcomes. Nonetheless, smaller models such as LLaMA 3 8B exhibit significant sensitivity to few-shot prompting. While it attains a zero-shot macro-F1 score of 0.63, performance significantly declines to 0.44 in the 1-shot scenario, with only a modest recovery to 0.63 at the 5-shot level.

In summary, lexicon-based sentiment analysis methods like VADER and TextBlob are insufficient for TBFSA because they fail to capture contextual financial semantics. Discriminative transformer-based models such as DistilFinRoBERTa, FinBERT, and FinBERT-Tone provide quantifiable enhancements but remain inadequate regarding precision and entity-level interpretability. Domain-adapted models like DeBERTa-v3-absa-v1.1, although tailored for target/aspect-based tasks, are surpassed by generative LLMs such as ChatGPT variants and DeepSeek-R1.

The consistent success of ChatGPT-4, ChatGPT-4o, ChatGPT-o1, and DeepSeek-R1 on the TBFSA task demonstrates the efficacy of comprehensive pre-training, which equips these LLMs to perform exceptionally in zero/few-shot scenarios and generalize across several domains without requiring task-specific fine-tuning. Their consistent superiority over conventional lexicon-based systems and discriminative transformer-based models underscores a significant transition towards generative LLMs that integrate high adaptability with robust domain-agnostic generalization, thus providing an efficient substitute for resource-intensive supervised methods in specialized tasks such as TBFSA. Such granular, entity-specific sentiment interpretation holds substantial implications for investors, financial analysts, and algorithmic trading systems. These advanced models allow stakeholders to participate in more informed decision-making, potentially improving portfolio management techniques and optimizing market timing decisions.

However, implementing LLMs in financial markets presents obstacles. Significant processing complexity and inference latency limit their applicability in ultra-high-frequency trading, where execution times are quantified in milliseconds. Moreover, regulatory issues arise from the intrinsic opacity of LLM decision-making, which contradicts compliance requirements such as MiFID II and SEC Rule 15c3-5 that necessitate model interpretability for audit and risk governance. These limitations underscore the need for transdisciplinary innovation. The effective incorporation of LLMs into financial analytics will likely rely on hybrid architectures that combine language capabilities with conventional econometric models. These hybrid architectures hold the potential to revolutionize financial analytics, balancing traditional financial metrics' interpretability with AI's adaptive learning capabilities, and thereby mitigating the risks linked to opaque algorithmic decision-making. Resolving these complexities necessitates collaboration among AI researchers, economists, and regulatory authorities to ensure that innovations, such as federated learning for data privacy and synthetic financial text generation for enhanced training robustness, are implemented ethically and effectively.

5. Conclusions

This study offers a comprehensive evaluation of target-based financial sentiment analysis (TBFSA) by systematically comparing the effectiveness of cutting-edge generative large language models (LLMs)—including ChatGPT, DeepSeek, LLaMA, and Gemma—with conventional lexicon-based methods (VADER, TextBlob) and discriminative transformer-based models (FinBERT, DistilFinRoBERTa, FinBERT-Tone, and DeBERTa-v3-base-ABSA-v1.1).

The findings indicate that LLMs—especially ChatGPT variants (notably ChatGPT-o1) and DeepSeek-R1—surpass all baseline models in target-level sentiment analysis. Their capacity to deduce implicit sentiment, adapt to financial terminology, and function efficiently without task-specific fine-tuning makes them scalable, ready-to-deploy solutions for practical applications like algorithmic trading and real-time risk assessment. These findings bear immediate implications for financial institutions, fintech developers, and analysts seeking to incorporate sentiment-driven insights into investing and risk management processes.

Despite the promising findings, the study acknowledges numerous limitations. The investigation is confined to news articles from four prominent technological firms—Alphabet, Amazon, Netflix, and Nvidia—potentially constraining the generalizability of the findings to other industries or smaller market-cap companies with possibly distinct sentiment patterns. Furthermore, the study encompasses a short time frame (Sep 4, 2023, to Jan 30, 2024), offering short-term insights while potentially neglecting long-term patterns, seasonal fluctuations, and macroeconomic changes. In addition, the sole dependence on news articles neglects other vital data sources, such as social media sentiment, earnings reports, and macroeconomic indicators, which could enhance the

research. To address these constraints, future research could broaden the analysis to encompass various sectors and global markets, integrate additional data sources, and prolong the study over several years to assess LLM performance over various market regimes, including bulls and bear cycles. Moreover, enhancing prompt designs via automated techniques, investigating time-lagged sentiment effects, and improving the interpretability of LLM outputs signify promising avenues for attaining more robust, comprehensible, and sector-agnostic applications of LLM-driven financial sentiment research.

Data Availability

The dataset developed with this research is available at https://github.com/iftikharm895/Target-Based_Sentiment_Analysis_in_Financial_News. Due to copyright constraints, only URLs with manual annotations are publicly released, with full news content accessible through a Bloomberg Terminal.

References

[1] M. Wu, G. Subramaniam, Z. Li, X. Gao, Using AI Technology to Enhance Data-Driven Decision-Making in the Financial Sector, in: Artificial Intelligence-Enabled Businesses: How to Develop Strategies for Innovation, 2025, pp. 187–207.

[2] Y. Yang, Y. Zhang, M. Wu, K. Zhang, Y. Zhang, H. Yu, Y. Hu, B. Wang, TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets, arXiv preprint arXiv:2502.01506 (2025).

[3] E. Cambria, B. White, Jumping NLP curves: A review of natural language processing research, IEEE Computational intelligence magazine 9 (2024) 48–57.

[4] K. Du, F. Xing, R. Mao, E. Cambria, Financial sentiment analysis: Techniques and applications, ACM Computing Surveys 56 (2024) 1–42.

[5] M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik, D. Trajanov, Sentiment analysis in finance: From transformers back to explainable lexicons (xlex), IEEE Access 12 (2024) 7170–7198.

[6] A. Matarazzo, R. Torlone, A Survey on Large Language Models with some Insights on their Capabilities and Limitations, arXiv preprint arXiv:2501.04040 (2025).

[7] R. Wadawadagi, S. Tiwari, V. Pagi, Polarity-aware deep attention network for aspect-based sentiment analysis, Progress in Artificial Intelligence 14 (2025) 33–48.

[8] S. Deng, Y. Zhu, Y. Yu, X. Huang, An integrated approach of ensemble learning methods for stock index prediction using investor sentiments, Expert Systems with Applications 238 (2024) 121710.

[9] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, P. Takala, Good debt or bad debt: Detecting semantic orientations in economic texts, Journal of the Association for Information Science and Technology 65 (2014) 782–796.

[10] F. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, F. Benevenuto, Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Science 5 (2014) 1–29.

[11] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.

[12] W. Aljedaani, F. Rustam, M. Mkaouer, A. Ghallab, V. Rupapara, P. Washington, E. Lee, I. Ashraf, Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry, Knowledge-Based Systems 255 (2022) 109780.

[13] M. Siek, V. Chandra, Analysis of News Sentiment for Stock Price Prediction Using VADER Sentiment, in: 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS), IEEE, 2024, pp. 1–6.

[14] T. Saleem, U. Yaqub, S. Zaman, Twitter sentiment analysis and bitcoin price forecasting: implications for financial risk management, The Journal of Risk Finance 25 (2024) 407–421.

[15] B. Nagendra, S. Chandar, J. Simha, J. Bazil, Financial Lexicon based Sentiment Prediction for Earnings Call Transcripts for Market Intelligence, in: 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN), IEEE, 2024, pp. 595–603.

[16] V. Khandelwal, H. Varshney, G. Munjal, Sentiment analysis-based stock price prediction using machine learning, in: 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), IEEE, 2024, pp. 182–187.

[17] I. Muhammad, M. Rospocher, On Assessing the Performance of LLMs for Target-Level Sentiment Analysis in Financial News Headlines, Algorithms 18 (2025) 46.

[18] J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[19] Z. Liu, D. Huang, K. Huang, Z. Li, J. Zhao, Fin-

bert: 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, 2021, pp. 4513–4519.

[20] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063 (2019).

[21] M. Mahendran, A. Gokul, P. Lakshmi, S. Pavithra, Comparative Advances in Financial Sentiment Analysis: A Review of BERT, FinBert, and Large Language Models, in: 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, 2025, pp. 39–45.

[22] A. Huang, H. Wang, Y. Yang, FinBERT: A large language model for extracting information from financial text, Contemporary Accounting Research 40 (2023) 806–841.

[23] A. Atak, Exploring the sentiment in Borsa Istanbul with deep learning, Borsa Istanbul Review 23 (2023) S84–S95.

[24] Y. Shen, P. Zhang, Financial sentiment analysis on news and reports using large language models and finbert, in: 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), IEEE, 2024, pp. 717–721.

[25] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, Bloomberggpt: A large language model for finance, arXiv preprint arXiv:2303.17564 (2023).

[26] Z. Wang, Y. Li, J. Wu, J. Soon, X. Zhang, Finvis-gpt: A multimodal large language model for financial chart analysis, arXiv preprint arXiv:2308.01430 (2023).

[27] Y. Yang, Y. Tang, K. Tam, Investlm: A large language model for investment using financial domain instruction tuning, arXiv preprint arXiv:2309.13064 (2023).

[28] P. Agarwal, A. Gupta, Strategic business insights through enhanced financial sentiment analysis: A fine-tuned llama 2 approach, in: 2024 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2024, pp. 1446–1453.

[29] W. Kang, X. Yuan, X. Zhang, Y. Chen, J. Li, ChatGPT-based Sentiment Analysis and Risk Prediction in the Bitcoin Market, Procedia Computer Science 242 (2024) 211–218.

[30] K. Krippendorff, Content analysis: An introduction to its methodology, Sage publications, 2018.

[31] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Computational linguistics 34 (2008) 555–596.

[32] S. Sohangir, N. Petty, D. Wang, Financial sentiment lexicon analysis, in: 2018 IEEE 12th international conference on semantic computing (ICSC), IEEE, 2018, pp. 286–289.

[33] L. Nemes, A. Kiss, Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication 5 (2021) 375–394.

[34] M. El Idrissi, N. Chafik, R. Tachicart, Stock Price Prediction Using Sentiment Analysis and LSTM Networks, in: IBIMA Conference on Artificial intelligence and Machine Learning, Springer Nature Switzerland, 2024, pp. 149–156.

[35] A. Patil, H. Sharma, A. Sinha, Sentiment Analysis of Financial News and its Impact on the Stock Market, in: 2024 2nd World Conference on Communication & Computing (WCONF), IEEE, 2024, pp. 1–5.

[36] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).

[37] H. Yang, C. Zhang, K. Li, Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis, in: Proceedings of the 32nd ACM international conference on information and knowledge management, 2023, pp. 5117–5122.

[38] F. Voigt, J. Calero, K. Dahal, Q. Wang, K. V. Luck, P. Stelldinger, Towards machine learning based text categorization in the financial domain, in: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS), IEEE, 2024, pp. 1–6.

[39] A. Dmonte, E. Ko, M. Zampieri, An Evaluation of Large Language Models in Financial Sentiment Analysis, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 4869–4874.

[40] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Transactions on Knowledge Discovery from Data 18 (2024) 1–32.

[41] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC genomics 21 (2020) 1–13.

[42] M. Rospocher, S. Eksir, Assessing fine-grained explicitness of song lyrics, Information 14 (2023).