algorithms

MDPI

Article

On Assessing the Performance of LLMs for Target-Level Sentiment Analysis in Financial News Headlines

Iftikhar Muhammad [1] and Marco Rospocher [2,*]

[1] Department of Economics, University of Verona, 37129 Verona, Italy; iftikhar.muhammad@univr.it
[2] Department of Foreign Languages and Literatures, University of Verona, 37129 Verona, Italy
* Correspondence: marco.rospocher@univr.it

Abstract: The importance of sentiment analysis in the rapidly evolving financial markets is widely recognized for its ability to interpret market trends and inform investment decisions. This study delves into the target-level financial sentiment analysis (TLFSA) of news headlines related to stock. The study compares the performance in the TLFSA task of various sentiment analysis techniques, including rule-based models (VADER), fine-tuned transformer-based models (DistilFinRoBERTa and Deberta-v3-base-absa-v1.1) as well as zero-shot large language models (ChatGPT and Gemini). The dataset utilized for this analysis, a novel contribution of this research, comprises 1476 manually annotated Bloomberg headlines and is made publicly available (due to copyright restrictions, only the URLs of Bloomberg headlines with the manual annotations are provided; however, these URLs can be used with a Bloomberg terminal to reconstruct the complete dataset) to encourage future research on this subject. The results indicate that the fine-tuned Deberta-v3-base-absa-v1.1 model performs better across all evaluation metrics than other evaluated models in TLFSA. However, LLMs such as ChatGPT-4, ChatGPT-4o, and Gemini 1.5 Pro provide similar performance levels without the need for task-specific fine-tuning or additional training. The study contributes to assessing the performance of LLMs for financial sentiment analysis, providing useful insights into their possible application in the financial domain.

Keywords: financial sector; target-level sentiment analysis; traditional methods; large language models; zero-shot learning

Academic Editors: Yuyi Mao, Jiawei Shao and Frank Werner

Received: 10 December 2024
Revised: 23 December 2024
Accepted: 10 January 2025
Published: 13 January 2025

Citation: Muhammad, I.; Rospocher, M. On Assessing the Performance of LLMs for Target-Level Sentiment Analysis in Financial News Headlines. *Algorithms* **2025**, *18*, 46. https://doi.org/10.3390/a18010046

1. Introduction

The financial sector, known for its dynamic nature, is greatly impacted by the continuous flow of news and information that changes market dynamics and investor views [1]. Every day, a large number of articles and microblogs containing detailed financial information about traded corporations are distributed online. This generates a substantial quantity of disorganized written information, posing a significant difficulty for companies, investors, and analysts involved in data analysis. The conventional approaches of manual data analysis are time consuming and increasingly unfeasible due to the vast amount of data [2].

Automated sentiment analysis, a popular natural language processing (NLP) task, is a viable solution that enables the automatic identification of emotional tones in texts, thereby offering valuable insights into investor behavior and market trends [3,4]. The decision-making processes of investors and market stakeholders are significantly impacted by the sentiments conveyed through financial news [5]. For example, positive news regarding

a company's performance may increase investor confidence and increase stock prices, whereas negative news may incite dread and trigger selloffs.

Sentiment analysis in this field utilizes several strategies, such as lexicon-based approaches, conventional machine learning techniques, advanced deep learning models, and transformer-based architectures [6]. Large language models (LLMs) like ChatGPT and Gemini have the potential to greatly impact sentiment analysis, as they can effectively work in situations where there are little or no training data available without requiring substantial manual annotation effort for supervised training [7]. Rigorously evaluating their performance in sentiment analysis for specific domains, such as the financial sector, is essential to assess their suitability for integration into domain-specific tools, including financial reporting software.

Sentiment analysis can be performed at four different levels of detail: document, sentence, target, and aspect level. The document level evaluates the overall sentiment of the entire text, while the sentence level analyzes sentiment in individual sentences. The target level assesses sentiment toward specific entities or categories, referred to as targets within the text [8–10], and the aspect level focuses on sentiment toward different features or attributes of an entity. Prior research on financial sentiment analysis has predominantly focused on the document and sentence levels [11–13]. Nevertheless, target-based analysis is essential for accurately interpreting and responding to the complex sentiments of financial news, as these documents frequently contain multiple sentiments associated with distinct entities [4]. For example, a headline such as "Netflix shares dip despite Apple's surge in streaming subscribers" illustrates divergent sentiments associated with specific companies, Netflix and Apple, within a single sentence, emphasizing the necessity of nuanced, target-based sentiment analysis.

Despite the financial sector's significant relevance in target-level sentiment analysis, this domain is primarily underexplored with limited exceptions such as implementations in economic and political news [14] and several commercial products [15]. Additionally, the considerable interest in the integration of LLMs in various disciplines has prompted extensive investigation into their potential applications in a wide range of domains [16,17]. However, their deployment within the stock market, particularly in investigating their effectiveness in conducting target-level financial sentiment analysis in zero-shot learning scenarios, has not yet been investigated. Furthermore, there is a significant lack of comparative performance analyses between traditional models and LLMs for financial sentiment analysis tasks.

The objective of this research is to fill these gaps by assessing the capability of LLMs to analyze sentiment at the target level in the financial field with a specific focus on news related to the stock market. We believe that such an assessment is both timely and relevant with the potential to influence the development of professional financial reporting software. We compare the effectiveness of LLMs, such as ChatGPT and Gemini, to traditional models, such as VADER, as well as to fine-tuned transformer-based models, including DistilFinRoBERTa and Deberta-v3-base-absa-v1.1, to identify and categorize sentiment at the target level within financial news headlines. More specifically, we address the following research questions:

(RQ1) How effectively do zero-shot LLMs perform in target-level financial sentiment analysis, and how do they compare to other state-of-the-art sentiment analysis techniques?

(RQ2) How do the performances of different versions of ChatGPT (specifically, ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o) compare to those of Gemini models (including Gemini 1 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro) in target-level sentiment analysis tasks under zero-shot scenarios?

The work detailed in this paper makes the following novel key contributions compared to existing studies:

(1) We present and publicly release a novel dataset that is specifically manually annotated for target-level sentiment analysis. This dataset consists of 1476 headlines from the financial sector. Studying target-level sentiment in news headlines is crucial because headlines shape immediate perceptions, influence public opinion, and dominate social media sharing, offering a concise and impactful representation of the news that often carries more bias and emotional weight than the full text. This resource is a critical instrument for the development and evaluation of sentiment analysis models that differentiate and evaluate sentiments directed at specific entities within the financial domain, as it uniquely addresses the nuanced requirements of target-level analysis in financial texts. To the best of our knowledge, no other dataset with the given characteristics is currently available;

(2) Utilizing the contributed dataset, we present a thorough comparative analysis of LLMs compared to conventional sentiment analysis methods, emphasizing their strengths and weaknesses in the context of target-level financial sentiment analysis. This study is one of the first to systematically assess the efficacy of LLMs, such as different versions of Gemini and ChatGPT, in conducting target-level financial sentiment analysis within a zero-shot learning approach. Our work thus contributes to the development of NLP tools in the financial sector, enabling stakeholders to make more informed decisions by utilizing target-level sentiment analysis.

The paper is organized as follows: Section 2 reviews related literature. Section 3 delineates the empirical methodology and the dataset creation and annotation process. The findings and discussion are reported in Section 4. Finally, Section 5 concludes the study and offers possible future research directions in the field of financial sentiment analysis.

2. Literature Review

2.1. Financial Sentiment Analysis

Financial sentiment analysis is a distinct area of study in data analytics that specifically deals with the interpretation of emotions, opinions, and attitudes conveyed in financial texts [18]. These encompass many types of information, such as news stories, social media posts, analyst reports, and earnings call transcripts. The main objective is to assess the sentiment toward financial markets, specific companies, or broader economic conditions, as these factors significantly impact investment choices and market dynamics [19].

The theory that public sentiment affects financial markets was initially introduced by [20], who attributed market behaviors to "animal spirits"—alternating waves of optimism and pessimism. This concept has been substantiated by a multitude of studies that have shown substantial correlations between investor sentiment and market fluctuations in asset prices and trading volumes [21,22].

Ref. [22] made significant contributions to the field of sentiment analysis for financial texts utilizing lexicon-based techniques, including the Bag of Words (BoW) model, to identify underlying sentiments in financial news. This method entails creating word lists with each word labeled as either having a positive or negative sentiment. Ref. [23] created a financial emotion dictionary specifically designed for evaluating 10-K reports, emphasizing the need to have lexicons that are tailored to specific sectors. Although lexicon-based analysis is widely used, it has been criticized for its simplistic approach and lack of consideration for word context [24–27].

In order to overcome these constraints, researchers have transitioned to employing machine learning methodologies. Machine learning frameworks, which are trained on labeled financial texts, offer more accurate sentiment assessments by adapting to word

context, unlike static lexicon-based algorithms [28]. Substantial advancements have been achieved in the field, such as the introduction of the FinancialPhraseBank dataset and the utilization of machine learning for sentiment detection [29]. Additionally, Ref. [30] suggested a competitive approach that relies on semantic characteristics.

The field has undergone further development following the integration of deep learning methods into natural language processing (NLP), resulting in considerable improvements in sentiment analysis capabilities. Transformer-based models such as BERT, presented by [31], have shown outstanding performance in sentiment classification tasks. This is attributed to their capacity to understand context, capture long-term relationships, and decrease data complexity [32]. Furthermore, the implementation of FinBERT, which is a specialized version of BERT [33] for the financial domain, has enhanced the reliability and accuracy of financial sentiment analysis [34,35].

The latest advancements in LLMs have significantly improved their effectiveness for NLP. New standards have been established by innovative techniques such as instruction tuning [17] and the retrieval-augmented framework for strengthening sentiment analysis [36]. The models ChatGPT, Gemini, and LLaMAs are presently under evaluation to determine their ability to forecast market trends based on news headlines [37,38]. This suggests a positive outlook for LLMs in the field of financial analytics.

2.2. Evolution of Target and Aspect-Based Sentiment Analysis

The primary focus of conventional sentiment analysis was the assessment of the general sentiment at the sentence or document level. Nevertheless, this method was insufficient in capturing subtle opinions regarding particular elements or entities. In order to overcome this constraint, researchers have developed target-based and aspect-based sentiment analysis techniques. These approaches seek to offer a more detailed comprehension of sentiment specifically connected to particular aspects or entities [14]. Target-level sentiment analysis (TLSA) is a method used to determine the general sentiment linked to a specific entity, which is known as the target. Conversely, aspect-based sentiment analysis (ABSA) concentrates on several facets of an item, such as its quality, robustness, or pricing [39].

The techniques used in target-level and aspect-based sentiment analysis vary from conventional machine learning methods to advanced deep learning and transformer-based models, which have greatly improved accuracy and contextual comprehension. Refs. [40–42] employed conventional methods that involved the use of algorithms such as Naive Bayes, Logistic Regression, and Support Vector Machines. These algorithms were used for extracting features and predicting sentiment in relation to specified entities or aspects.

The efficacy of TLSA and ABSA has been further enhanced by the emergence of deep learning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which include Long Short-Term Memory (LSTM) networks and Gated Recurrent Units, have been instrumental in capturing intricate patterns and contextual information. The use of CNNs and LSTMs enhanced by attention mechanisms to concentrate more accurately on relevant text segments when assessing sentiment for specific aspects is highlighted in studies by [43,44].

Transformer-based models, specifically BERT, have significantly transformed the fields of TLSA and ABSA by offering an unparalleled level of comprehension in terms of context and semantics [45,46]. These models achieve superior sentiment assessment performance by utilizing extensive pre-training on large datasets, which is followed by task-specific fine-tuning.

Although BERT-based models are effective, they encounter problems such as the need for pre-training that is appropriate to a particular domain and the management of ambiguous contexts [47]. In order to address these constraints, large language models (LLMs), including LLaMa, ChatGPT, and Gemini, have been implemented for TLSA and ABSA tasks [7]. Ref. [48] conducted experiments with the SemEval-2016 for the Restaurant and Laptop domains to investigate GPT's capabilities for the Aspect Sentiment Triplet Extraction (ASTE) subtask. In a similar manner, Ref. [49] employed GPT to forecast sentiment scores for different elements derived from bakery reviews. Nevertheless, there is a scarcity of thorough comparison assessments on different LLMs in specific fields like finance.

Building on this literature review, it is clear that preliminary investigations into the application of LLMs for ABSA and TLSA demonstrate potential; however, the specificity and complexity of domains such as finance necessitate a more comprehensive examination. Our research represents a significant effort to assess the efficacy of LLMs for TLSA in the financial sector. By conducting a comparative analysis of different LLMs, specifically ChatGPT versions 3.5, 4, and 4o, as well as the Gemini series—Gemini 1 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro—in conjunction with traditional methods like VADER, DistilFinRoBERTa, and DeBERTa-v3-base-ABSA-v1.1, our study assess their capabilities in the complex context of financial sentiment analysis.

3. Methodology

This section outlines the methodology used to evaluate LLMs like ChatGPT and Gemini, as well as traditional financial sentiment analysis methods such as VADER, DistilFinRoBERTa, and DeBERTa-v3-base-ABSA-V1.1, in the context of TLFSA. We elaborate on our methods for collecting data and the use of well-established methodologies for analyzing financial sentiment. Afterward, we fine-tune the DeBERTa-v3-base-ABSA-V1.1 model using our curated dataset. Subsequently, we formulate a precise prompt tailored for the ChatGPT and Gemini models to assess the effectiveness of these LLMs in TLFSA tasks. In the end, a thorough examination is carried out to compare traditional approaches with LLMs in the task of TLFSA.

3.1. Dataset Creation and Annotation

To conduct a comprehensive examination in this research, news headlines pertaining to well-known stock companies, namely Amazon, Netflix, Nvidia, and Alphabet, were collected. These headlines were sourced from Bloomberg's financial news service, Bloomberg Terminal. Bloomberg is a well-acknowledged platform that is extensively utilized for consolidating financial news data from many sources [50–52]. The dataset encompasses the time period from 1 September 2023, to 31 January 2024.

The dataset consists of 1476 distinct news headlines, which are each linked to a particular stock company (referred to as the "target"), a timestamp, and a URL. Keywords relevant to the targeted companies, such as "target-company" and "target-company in headlines", were used during the data-collecting phase from the Bloomberg Terminal. For example, when specifically analyzing Amazon.com, Inc. as the subject company, keywords such as 'Amazon.com Inc.' and 'Amazon in headlines' were employed to aid in extracting pertinent articles. As a result, this method produced a collection of news articles that were pertinent to the pre-identified target companies. These articles had headlines that mentioned these companies. Subsequently, the gathered articles underwent processing to extract crucial information such as headlines, timestamps, and URLs, which were later recorded in a database, as seen in Figure 1.

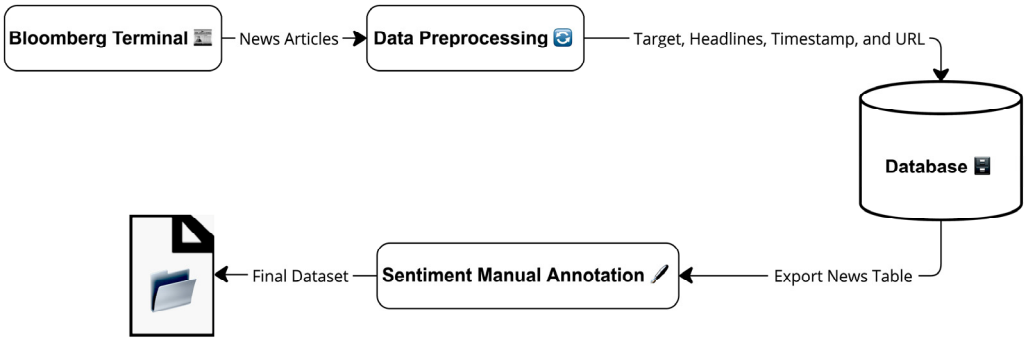

Figure 1. Dataset creation workflow.

To determine the dataset's dependability and confirm the experimental evaluation, each headline is manually annotated to evaluate the sentiment toward the considered target company. This technique involves examining the sentiment expressed in the headlines in relation to the set targets and categorizing these sentiments as positive, negative, or neutral. The annotations are performed by three annotators, experts in finance and economics, all having a high level of English proficiency (CEFR level C1).

The annotators are given comprehensive instructions that outline the tasks required for target-level annotation. Here are some examples that demonstrate the annotation guidelines given to the annotators (the full version of the guidelines is released together with the dataset):

- Assess the sentiment of each headline exclusively with respect to the target rather than considering the overall sentiment of the entire headline. Annotators should assume the viewpoint of an investor and evaluate the probable influence of the headline on investment decisions.
- Assign a sentiment score to a headline in relation to the target, using the following scale: (0) for a neutral sentiment, (−1) for a negative sentiment, and (+1) for a positive sentiment.
- Formulate annotations based solely on the sentiment directly conveyed in the headline rather than relying on external information.
- Assign a neutral sentiment score of 0 to headlines that contain muddled sentiments, uncertainty, or ambiguity.
- Assign a neutral sentiment when there is no evident positive or negative sentiment regarding the target.

The manual annotation technique for news headlines involves an initial phase when the three annotators individually examine the same subset of 150 headlines. To quantify the level of agreement among the three coders and ensure the consistency of the manual labeling process, we employ Krippendorff's Alpha [53] as a measure of intercoder agreement. The calculated Krippendorff's Alpha score is 0.8188, which indicates a satisfactory level of agreement among the coders, thus confirming the reliability of the contributed annotations.

After ensuring the satisfactory level of intercoder agreement, the remaining headlines, which amount to 1326, are divided equally among the three raters. Each rater is assigned the task of personally annotating 442 headlines. Table 1 presents examples of annotated headlines.

Moreover, Figure 2 shows the distribution of sentiment specific to each target. Overall, the dataset is slightly unbalanced with roughly 49% of positive headlines, 28% of negative headlines, and 23% of neutral headlines.

Table 1. Examples of annotated headlines.

Target	Headline	Label
Amazon	Pioneer Disciplined Growth Buys More Nike Class B, Cuts Amazon	−1
Amazon	Amazon to Invest Over $440 m to Boost Delivery Drivers' Pay	1
Netflix	Netflix, Tarantino Win $20 Million California Film Tax Credits	1
Netflix	Netflix Shares Dip Despite Apple's Surge in Streaming Subscribers	−1
Netflix	Netflix CFO Speaks at Bank of America Conference	0
Nvidia	Nvidia to Partner with India's Tata, Reliance for AI Development	1
Nvidia	Neuberger Berman Guardian Buys More Walmart, Cuts Nvidia	−1
Alphabet	Alphabet Hits 52-Week High at $139.17	1

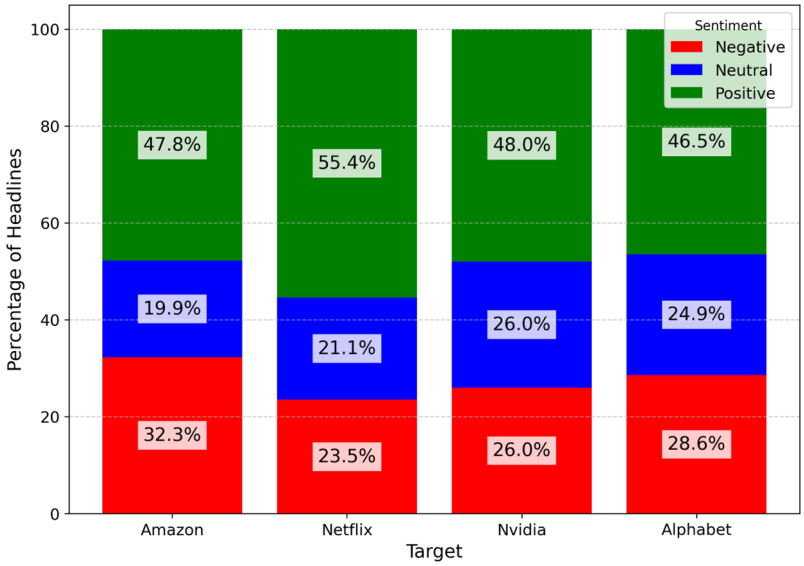

Figure 2. A visual representation of sentiment distribution of news headlines for the targets in the dataset.

Table 2 displays quantitative measures pertaining to our dataset, including the overall count of articles, as well as the standard deviation and average of daily articles, which are categorized by each target.

Table 2. Dataset statistics.

Target	Articles	Daily Articles
Amazon	573	5.07 (4.41)
Netflix	298	3.77 (5.14)
Nvidia	392	3.70 (4.28)
Alphabet	213	3.04 (3.95)
Total	1476	15.58 (17.78)

In order to enhance the research community and guarantee the reproducibility of our procedures, we publicly release our curated dataset (due to copyright restrictions, this dataset only includes URLs of headlines; the actual text of the headlines cannot be released. However, the URLs can be used by users with access to a Bloomberg terminal to recreate

the dataset. The dataset is available at: https://github.com/iftikharm895/Target-Level_Financial_Sentiment_Analysis—accessed on 23 December 2024). This dataset, consisting of news headlines related to prominent stock companies and their relevant sentiments, is a substantial resource and is anticipated to be highly beneficial for researchers and practitioners who specialize in utilizing machine learning methods for target-level financial sentiment analysis. This work is expected to foster additional study in this field, thereby enhancing our collective comprehension of the complex connections between market behavior, AI models, and financial news sentiment.

3.2. Target-Level Sentiment Classification Using Traditional Methods

This study utilizes three conventional models for financial sentiment analysis: VADER, DistilFinRoBERTa, and Deberta-v3-base-absa-v1.1. The VADER (https://github.com/cjhutto/vaderSentiment—accessed on 23 December 2024) tool, developed by [54], has been widely used by researchers in the field of financial sentiment analysis. It is a rule-based tool for performing document/sentence sentiment analysis. While specifically attuned to social media, VADER also performs well in other domains. Some notable studies that have utilized VADER in the financial domain include [2,24,55–57].

DistilFinRoBERTa (https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis—accessed on 23 December 2024) is a lightweight variant of FinBERT, which is a model that was specifically pre-trained for understanding financial texts. This entails utilizing a financial collection of texts and refining the model for categorizing financial sentiment, incorporating resources like the in-house documents and annotations from John-Snow LAB and the Financial PhraseBank by [29]. DistilFinRoBERTa has been implemented in the financial sector for document/sentence sentiment analysis by scholars such as [58,59].

DeBERTa-v3-base-ABSA-v1.1 (https://huggingface.co/yangheng/deberta-v3-base-absa-v1.1—accessed on 23 December 2024) is a model specifically developed for ABSA. It uses the FAST-LCF-BERT model, which includes Microsoft/DeBERTa-v3-base and is sourced from PyABSA [60,61]. This model is trained using a large dataset of over 30,000 ABSA samples. It is then fine-tuned with an additional 180,000 examples from various ABSA datasets (https://github.com/yangheng95/ABSADatasets—accessed on 23 December 2024), including SemEval-2014 Task 4 (Laptop14 and Restaurant14) [62], MAMS [63], SemEval-2016 Task 5 (Restaurant-16) [64], and other datasets specific to product categories such as T-shirts [65,66] and televisions [67]. This model has been utilized by various researchers for ABSA tasks, such as [68–70]. We employ DeBERTa-v3-base-ABSA-v1.1 in two configurations: the pre-trained model as provided and a fine-tuned version developed through 5-fold cross-validation (5-fold cross-validation tests the model on the entire dataset by ensuring every data point is used for both training and testing but never at the same time. This helps avoid overfitting and provides a more reliable measure of model performance). For fine tuning, the model was trained on each fold for 10 epochs using default parameters (‘model’: <class ‘pyabsa.tasks.AspectPolarityClassification.models.__lcf__.fast_lsa_t_v2.FAST_LSA_T_V2’>, ‘optimizer’: ‘adamw’, ‘learning_rate’: 2e-05, ‘cache_dataset’: True, ‘warmup_step’: −1, ‘deep_ensemble’: False, ‘use_bert_spc’: True, ‘max_seq_len’: 80, ‘patience’: 99999, ‘SRD’: 3, ‘dlcf_a’: 2, ‘dca_p’: 1, ‘dca_layer’: 3, ‘use_syntax_based_SRD’: False, ‘sigma’: 0.3, ‘lcf’: ‘cdw’, ‘lsa’: False, ‘window’: ‘lr’, ‘eta’: 1, ‘eta_lr’: 0.1, ‘dropout’: 0.5, ‘l2reg’: 1e-06, ‘num_epoch’: 10, ‘batch_size’: 16, ‘initializer’: ‘xavier_uniform_’, ‘seed’: 52, ‘output_dim’: 3, ‘log_step’: 5, ‘dynamic_truncate’: True, ‘srd_alignment’: True, ‘evaluate_begin’: 0, ‘similarity_threshold’: 1, ‘cross_validate_fold’: −1, ‘use_amp’: False, ‘overwrite_cache’: False, ‘pretrained_bert’: ‘yangheng/deberta-v3-base-absa-v1.1’). The entire 5-fold cross-validation process was completed in 8 min on a standard NVIDIA RTX 4090 GPU.

We remark that VADER and DistilFinRoBERTa provide sentence/document level sentiment analysis, while DeBERTa-v3-base-ABSA-v1.1 is developed to perform target/aspect-based sentiment analysis.

3.3. Target-Level Sentiment Classification with ChatGPT and Gemini

The recent rise of LLMs has been highly notable, showcasing their exceptional capacity to execute a wide range of text-oriented tasks [71]. Presently, OpenAI's ChatGPT (https://chatgpt.com/ (Versions available as of 30 June 2024)) and Google's Gemini (https://gemini.google.com/app (Versions available as of 30 June 2024)) are two highly esteemed and widely used LLMs. The current study utilized multiple versions of the ChatGPT and Gemini models, such as ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini 1 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro, to perform TLFSA. No fine tuning of the models was conducted, as the objective of this study was to evaluate the zero-shot performance of these advanced generative AI tools on the given task. Interactions with ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o were carried out using the ChatGPT user interface (UI) in its default configuration, accessible to all users, with a preset temperature of 0.7. Additionally, access to Gemini models (Gemini 1 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro) was obtained via the Gemini API. The recommended temperature settings provided by the API were followed: 0.9 for Gemini 1 Pro and 1.0 for both Gemini 1.5 Flash and Gemini 1.5 Pro.

To evaluate the performance of several ChatGPT and Gemini models in TLFSA, we utilized a zero-shot prompting method (we also explored the few-shot learning approach by adding a small number of gold-annotated headlines as examples to the prompt, aiming to guide the LLM in performing the task. However, this did not result in significant performance improvements. A more comprehensive evaluation of the few-shot learning capabilities of these models for this task is planned for future work) [72,73]; that is, the model is instructed with an indication of how to perform a given task without providing any gold annotated example (i.e., a headline and the corresponding manual annotation in our context). More in detail, the following prompt was crafted:

Evaluate the sentiment conveyed by the headline with respect to the {target} from an investment perspective. Employ a three-tier scale for this assessment: allocate a score of (−1) for a negative sentiment, (0) for a neutral sentiment, and (+1) for a positive sentiment. Assign a neutral score of 0 if the headline is vague or does not clearly express positive or negative implications about the {target}.

By utilizing zero-shot prompting, we exploit the pre-existing training and comprehension of language context and patterns in ChatGPT and Gemini to produce desirable responses without the need for additional training or task-specific fine tuning. This method enables a straightforward evaluation of the intrinsic capacities of these models in sentiment analysis without any alterations, showcasing their potential for effortless integration and utilization in different contexts.

The performance of all the above models and methods was evaluated using a variety of well-established metrics for sentiment/classification tasks [74–76], such as precision, recall, F1-measure, and accuracy. Specifically, we report the aggregated performance across the three classes (Positive, Neutral, Negative) using both macro-averaging, where metrics are calculated for each class separately and averaged equally across all classes, and weighted averaging, where metrics are weighted by the number of samples in each class. These metrics provide a comprehensive evaluation of each model's effectiveness in detecting target-level sentiments, even in the presence of a slightly imbalanced class distribution, as seen in our dataset [75,76].

4. Results and Discussion

Table 3 reports the scores obtained by each considered model on the whole novel dataset contributed with this work.

Table 3. Performance outcomes of target-level sentiment classification across models. Best scores are in bold.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Weighted Precision	Weighted Recall	Weighted F1-Score
VADER	0.4939	0.5093	0.4736	0.4753	0.5336	0.4939	0.5010
DistilFinRoBERTa	0.5129	0.6182	0.5357	0.5010	0.6757	0.5129	0.5283
DeBERTa-v3	0.5169	0.6736	0.5833	0.5367	0.7451	0.5169	0.5324
DeBERTa-v3 (fine-tuned)	**0.8679**	**0.8499**	**0.8414**	**0.8448**	**0.8657**	**0.8679**	**0.8662**
ChatGPT-3.5	0.7527	0.7716	0.7470	0.7385	0.8179	0.7527	0.7689
ChatGPT-4	0.8591	0.8395	0.8314	0.8337	0.8646	0.8591	0.8605
ChatGPT-4o	0.8354	0.8213	0.8296	0.8187	0.8613	0.8354	0.8429
Gemini 1 Pro	0.7791	0.7522	0.7293	0.7382	0.7747	0.7791	0.7743
Gemini 1.5 Pro	0.8266	0.8161	0.8253	0.8123	0.8570	0.8266	0.8350
Gemini 1.5 Flash	0.8313	0.8041	0.8031	0.8030	0.8276	0.8313	0.8290

VADER exhibits the lowest performance across all criteria, highlighting its constraints in target-level sentiment analysis. VADER demonstrates a macro precision of 0.5093, a macro F1-score of 0.4753, a macro recall of 0.4736, and an overall accuracy of 0.4939. The weighted metrics provide more evidence of its insufficiency with a weighted F1-score of 0.5010, a weighted precision of 0.5336, and a weighted recall of 0.4939. VADER, a lexicon-based technique originally designed for general sentiment analysis, particularly in social media contexts, faces difficulties in handling the subtle language and specialist terminology commonly found in financial literature. The system's static and rule-based nature hinders its ability to accurately understand the changing linguistic nuances commonly seen in financial discussions. This limitation restricts its usefulness in dynamic and contextually sensitive situations.

Pre-trained models such as DistilFinRoBERTa and DeBERTa-v3-base-ABSA-v1.1 exhibit some improvement compared to VADER. DistilFinRoBERTa demonstrates a marginal improvement in accuracy, reaching a value of 0.5129. This improvement is accompanied by enhanced macro measures, including a recall of 0.5357, precision of 0.6182, and F1-score of 0.5010. DeBERTa-v3-base-ABSA-v1.1 demonstrates improved performance, with an accuracy of 0.5169, macro recall of 0.5833, macro precision of 0.6736, and macro F1-score of 0.5367. These models perform better than VADER, which is likely because of their sophisticated structures and specialized training. DistilFinRoBERTa benefits from effective pre-training on various datasets in the financial domain, while DeBERTa-v3-base-ABSA-v1.1 uses a unique disentangled attention mechanism that is crucial for understanding target-level sentiments. Moreover, we recall that DeBERTa-v3-base-ABSA-v1.1 is specifically fine-tuned for target/aspect-level sentiment analysis, while VADER and DistilFinRoBERTa are meant for document/sentiment-level sentiment analysis.

The fine-tuned DeBERTa-v3-base-ABSA-v1.1 model stands out by achieving the highest recorded accuracy of 0.8679 in this investigation. This model underwent thorough training and fine-tuning using our financial dataset, employing a 5-fold cross-validation approach. The model attains a macro recall of 0.8414, macro precision of 0.8499, and macro F1-score of 0.8448. The weighted metrics also demonstrate its exceptional performance. The results underscore the significant advantages of refining this model on domain-specific data.

Among the LLMs, ChatGPT-4 achieves an outstanding performance, ranking second only to the highly optimized DeBERTa-v3-base-ABSA-v1.1. It achieves a macro aver-

age F1-score of 0.8337 and an accuracy of 0.8591. The performance of Gemini models, specifically Gemini 1.5 Pro and Gemini 1.5 Flash, is noteworthy, demonstrating significant improvements compared to previous versions such as Gemini 1 Pro.

To summarize, conventional models for document/sentence-level sentiment analysis, like VADER and DistilFinRoBERTa, proved inadequate for TLFSA, even if the latter is trained on financial text. This highlights the limitations of traditional sentiment analysis tools in handling the nuanced and context-specific requirements of target-level sentiment analysis, especially in specialized domains like finance. The pre-trained DeBERTa-v3-base-ABSA-v1.1, despite being built for the considered task, exhibits limitations when applied to the financial domain, which is likely due to the different domain of its training data. This finding underscores the importance of domain adaptation in achieving optimal performance.

Fine-tuning DeBERTa-v3-base-ABSA-v1.1 on financial data significantly enhances its efficacy, as evidenced by the high performance achieved through 5-fold cross-validation. The success of this fine-tuning process illustrates the critical role of tailored training for domain-specific applications. This result aligns with prior research suggesting that pre-trained models require domain-specific adaptation to handle specialized vocabulary, semantics, and contextual relationships effectively.

LLMs such as ChatGPT-4, ChatGPT-4.0, and Gemini 1.5 Pro demonstrate substantial potential for the task. Despite not being explicitly fine-tuned on financial sentiment data, these models demonstrate performance levels comparable to those of fine-tuned DeBERTa-v3-base-ABSA-v1.1. This seems to suggest that LLMs have an inherent ability to generalize across tasks and domains, which is likely due to their extensive pre-training on diverse and large-scale datasets. Their capacity to perform well in TLFSA without additional training makes them highly versatile and efficient alternatives for direct application in such contexts, reducing the need for resource-intensive manual annotation and fine-tuning processes. These outcomes are in line with the findings observed in different contexts and experiments, such as [7,48,49].

The current study on TLFSA reveals notable progress in sentiment analysis in the financial sector, specifically with the fine-tuned DeBERTa-v3-base-ABSA-v1.1 and LLMs like ChatGPT-4, ChatGPT-4o, and Gemini 1.5 Pro. These models exhibit very high effectiveness in capturing nuanced emotions, making them valuable instruments for businesses dependent on sentiment analysis for strategic decision making. These models can be incorporated into financial reporting and predictive systems through user-friendly APIs, allowing seamless integration with existing financial software. For financial technology firms, the integration of these models could result in more advanced financial reporting tools that combine sentiment data elicited from news headlines with other economic indicators—such as stock prices, trading volumes, interest rates, and macroeconomic trends—offering more comprehensive market predictions and supporting informed decision making. The versatility of LLMs like ChatGPT-4, which do not require extensive fine tuning for financial data, reduces implementation time and costs, making them an appealing choice for both large institutions and smaller financial firms.

We recognize that integrating these models into real-world applications poses several challenges. A key issue is the interpretability of model outputs, as LLMs function as black boxes, generating outputs without transparent reasoning processes. This is particularly critical in financial contexts, where professionals require clear and justifiable explanations for sentiment predictions. Notably, ongoing research in Explainable AI [77] aims to address this challenge. Additionally, the LLM models used in our work are maintained and controlled by external entities (e.g., OpenAI, Google), which frequently update them to enhance their capabilities. This dynamic necessitates regular performance evaluations to

ensure their efficacy remains consistent. For example, periodic benchmarking on datasets such as the one contributed by this work can help monitor and validate their reliability for financial applications.

5. Conclusions

This study has performed a thorough assessment of target-level financial sentiment analysis (TLFSA), specifically examining the efficacy of conventional sentiment analysis techniques, such as VADER, DistilFinRoBERTa, and DeBERTa-v3-base-ABSA-v1.1, in comparison to advanced LLMs like ChatGPT and Gemini, within the domain of the stock market.

The study utilized a zero-shot prompting strategy to investigate the ability of LLMs to identify target-specific sentiments. This was accomplished by utilizing a novel, rigorously constructed dataset of annotated stock-related news headlines. The dataset is made available to the public to promote further research and innovation (due to copyright restrictions, only the URLs of Bloomberg headlines with the manual annotations are provided; however, these URLs can be used with a Bloomberg terminal to reconstruct the complete dataset).

VADER and DistilFinRoBERTa exhibit reduced effectiveness for TLFSA, lacking the specialized capabilities to accurately detect and evaluate the sentiment toward specific targets within the text. Hence, their performance in sentiment analysis at the target level is lower than models specifically developed for this task, such as DeBERTa-v3-base-ABSA-v1.1. Indeed, the fine-tuned version of the DeBERTa-v3-base-ABSA-v1.1 model exhibits exceptional performance, attaining the top scores in all evaluation metrics. LLMs also achieve very competitive performance, with ChatGPT-4, ChatGPT-4.0, and Gemini 1.5 Pro demonstrating impressive abilities. These models obtained almost ideal outcomes without requiring any training at all, emphasizing their potential for direct practical use in TLFSA.

The findings obtained from this study are relevant for investors, financial analysts, and fintech companies, illustrating how both conventional approaches and advanced LLMs can improve decision-making processes. The existing performance leaderboard offers beneficial information for choosing models depending on individual performance requirements and the availability of resources. This study not only expands our understanding of advanced sentiment analysis tools but also lays the foundation for future advancements that could improve the accuracy of financial sentiment analysis, facilitating more strategic and informed decision making in financial markets.

Although the study has shown encouraging results, it also acknowledges limitations, specifically the limited range of data used, which could impact the applicability of our findings to a broader context. Future studies should (i) expand the dataset to include a broader range of financial documents and extend the data collection timeframe to better capture the diversity of market dynamics and (ii) incorporate additional existing datasets to evaluate the generalizability of the findings. Furthermore, it will be essential to evaluate the potential of other LLMs for financial sentiment analysis as generative AI technologies continue to evolve, including models such as BLOOM, LLaMA, and competitive small-scale models (e.g., ORCA, OLMO). A thorough assessment of their few-shot performance is also recommended to further corroborate our preliminary observations on this aspect.

Author Contributions: Conceptualization, I.M. and M.R.; Data curation, I.M. and M.R.; Methodology, M.R.; Resources, I.M. and M.R.; Software, I.M.; Validation, I.M. and M.R.; Writing—original draft, I.M.; Writing—review and editing, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset developed with this research is available at https://github.com/iftikharm895/Target-Level_Financial_Sentiment_Analysis—accessed on 23 December 2024.

Due to copyright restrictions, this dataset only includes URLs of headlines with manual annotations; the actual text of the headlines cannot be released. However, the URLs can be used by users with access to a Bloomberg terminal to recreate the whole dataset.

Acknowledgments: The authors gratefully acknowledge Francesca Rossi for her invaluable contribution in the early stage of the work and dedicate it to her memory.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Adhikari, S.; Thapa, S.; Naseem, U.; Lu, H.Y.; Bharathy, G.; Prasad, M. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Netw.* **2023**, *164*, 115–123. [CrossRef] [PubMed]
2. Agarwal, A. Sentiment analysis of financial news. In Proceedings of the 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 25–26 September 2020; pp. 312–315.
3. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
4. Deng, S.; Zhu, Y.; Yu, Y.; Huang, X. An integrated approach of ensemble learning methods for stock index prediction using investor sentiments. *Expert Syst. Appl.* **2024**, *238*, 121710. [CrossRef]
5. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]
6. Simmering, P.F.; Huoviala, P. Large language models for aspect-based sentiment analysis. *arXiv* **2023**, arXiv:2310.18025.
7. Brauwers, G.; Frasincar, F. A survey on aspect-based sentiment classification. *ACM Comput. Surv.* **2022**, *55*, 1–37. [CrossRef]
8. Phan, H.T.; Nguyen, N.T.; Hwang, D. Aspect-level sentiment analysis: A survey of graph convolutional network methods. *Inf. Fusion* **2023**, *91*, 149–172. [CrossRef]
9. Atandoh, P.; Zhang, F.; Adu-Gyamfi, D.; Atandoh, P.H.; Nuhoho, R.E. Integrated deep learning paradigm for document-based sentiment analysis. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101578. [CrossRef]
10. Shirsat, V.S.; Jagdale, R.S.; Deshmukh, S.N. Document-level sentiment analysis from news articles. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 17–18 August 2017; pp. 1–4.
11. Lutz, B.; Pröllochs, N.; Neumann, D. Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning. *arXiv* **2018**, arXiv:1901.00400.
12. Husejinović, A.; Mašetić, Z. Document-based sentiment analysis on financial texts. In *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies*; Springer Nature: Cham, Switzerland, 2023; pp. 251–262.
13. Du, K.; Xing, F.; Cambria, E. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Trans. Manag. Inf. Syst.* **2023**, *14*, 1–24. [CrossRef]
14. Žitnik, S.; Blagus, N.; Bajec, M. Target-level sentiment analysis for news articles. *Knowl.-Based Syst.* **2022**, *249*, 108939. [CrossRef]
15. Ho, S.Y.; Choi, K.W.S.; Yang, F.F. Harnessing aspect-based sentiment analysis: How are tweets associated with forecast accuracy? *J. Assoc. Inf. Syst.* **2019**, *20*, 2. [CrossRef]
16. Zhang, B.; Yang, H.; Liu, X.Y. Instruct-finGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv* **2023**, arXiv:2306.12659. [CrossRef]
17. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large language models: A survey. *arXiv* **2024**, arXiv:2402.06196.
18. Yekrangi, M.; Nikolov, N.S. Domain-specific sentiment analysis: An optimized deep learning approach for the financial markets. *IEEE Access* **2023**, *11*, 70248–70262. [CrossRef]
19. Usmani, S.; Shamsi, J.A. LSTM-based stock prediction using weighted and categorized financial news. *PLoS ONE* **2023**, *18*, e0282234. [CrossRef]
20. Keynes, J.M. The general theory of employment. *Q. J. Econ.* **1937**, *51*, 209–223. [CrossRef]
21. Baker, M.; Wurgler, J. Investor sentiment in the stock market. *J. Econ. Perspect.* **2007**, *21*, 129–151. [CrossRef]
22. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Financ.* **2007**, *62*, 1139–1168. [CrossRef]
23. Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
24. Sohangir, S.; Petty, N.; Wang, D. Financial sentiment lexicon analysis. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 286–289.

25. Bonta, V.; Kumaresh, N.; Janardhan, N. A comprehensive study on lexicon-based approaches for sentiment analysis. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 1–6. [CrossRef]
26. Taj, S.; Shaikh, B.B.; Meghji, A.F. Sentiment analysis of news articles: A lexicon-based approach. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 30–31 January 2019; pp. 1–5.
27. Shang, L.; Xi, H.; Hua, J.; Tang, H.; Zhou, J. A lexicon-enhanced collaborative network for targeted financial sentiment analysis. *Inf. Process. Manag.* **2023**, *60*, 103187. [CrossRef]
28. Schumaker, R.P.; Chen, H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Trans. Inf. Syst. (TOIS)* **2009**, *27*, 1–19. [CrossRef]
29. Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; Takala, P. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.* **2014**, *65*, 782–796. [CrossRef]
30. Dridi, A.; Atzeni, M.; Recupero, D.R. FineNews: Fine-grained semantic sentiment analysis on financial microblogs and news. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2199–2207. [CrossRef]
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
32. Refaeli, D.; Hajek, P. Detecting fake online reviews using fine-tuned BERT. In Proceedings of the 2021 5th International Conference on E-Business and Internet, Singapore, 15–17 October 2021; pp. 76–80.
33. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
34. Farimani, S.A.; Jahan, M.V.; Fard, A.M.; Tabbakh, S.R.K. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowl.-Based Syst.* **2022**, *247*, 108742. [CrossRef]
35. Leippold, M. Sentiment spin: Attacking financial sentiment with GPT-3. *Financ. Res. Lett.* **2023**, *55*, 103957. [CrossRef]
36. Du, K.; Xing, F.; Mao, R.; Cambria, E. An evaluation of reasoning capabilities of large language models in financial sentiment analysis. In Proceedings of the IEEE Conference on Artificial Intelligence (IEEE CAI), Singapore, 25–27 June 2024.
37. Chen, W.; Du, J.; Zhang, Z.; Zhuang, F.; He, Z. A hierarchical interactive network for joint span-based aspect-sentiment analysis. *arXiv* **2022**, arXiv:2208.11283.
38. Lopez-Lira, A.; Tang, Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv* **2023**, arXiv:2304.07619. [CrossRef]
39. Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 November 2003; pp. 427–434.
40. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. *Encycl. Mach. Learn.* **2010**, *15*, 713–714.
41. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
42. Varghese, R.; Jayasree, M. Aspect-based sentiment analysis using support vector machine classifier. In Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 22–25 August 2013; pp. 1581–1586.
43. Wang, X.; Li, F.; Zhang, Z.; Xu, G.; Zhang, J.; Sun, X. A unified position-aware convolutional neural network for aspect-based sentiment analysis. *Neurocomputing* **2021**, *450*, 91–103. [CrossRef]
44. Sadr, H.; Pedram, M.M.; Teshnehlab, M. Convolutional neural network equipped with attention mechanism and transfer learning for enhancing performance of sentiment analysis. *J. AI Data Min.* **2021**, *9*, 141–151.
45. Rietzler, A.; Stabinger, S.; Opitz, P.; Engl, S. Adapt or get left behind: Domain adaptation through BERT language model fine-tuning for aspect-target sentiment classification. *arXiv* **2019**, arXiv:1908.11860.
46. Karimi, A.; Rossi, L.; Prati, A. Adversarial training for aspect-based sentiment analysis with BERT. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8797–8803.
47. Silva, E.H.D.; Marcacini, R.M. Aspect-based sentiment analysis using BERT with disentangled attention. In Proceedings of the LatinX in AI (LXAI) Research Workshop at ICML, Virtual, 19 July 2021.
48. Chumakov, S.; Kovantsev, A.; Surikov, A. Generative approach to aspect-based sentiment analysis with GPT language models. *Procedia Comput. Sci.* **2023**, *229*, 284–293. [CrossRef]
49. Magdaleno, D.; Montes, M.; Estrada, B.; Ochoa-Zezzatti, A. A GPT-based approach for sentiment analysis and bakery rating prediction. In *Mexican International Conference on Artificial Intelligence*; Springer Nature: Cham, Switzerland, 2023; pp. 61–76.
50. Chen, Q. Stock movement prediction with financial news using contextualized embedding from BERT. *arXiv* **2021**, arXiv:2107.08721.
51. Fedyk, A. Front-Page News: The effect of news positioning on financial markets. *J. Financ.* **2024**, *79*, 5–33. [CrossRef]

52. Fedyk, A.; Hodson, J. When can the market identify old news? *J. Financ. Econ.* **2023**, *149*, 92–113. [CrossRef]
53. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; Sage Publications: London, UK, 2018.
54. Hutto, C.; Gilbert, E. 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, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
55. Nemes, L.; Kiss, A. Prediction of stock value changes using sentiment analysis of stock news headlines. *J. Inf. Telecommun.* **2021**, *5*, 375–394. [CrossRef]
56. Padmanayana, V.; Bhavya, K. Stock market prediction using Twitter sentiment analysis. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2021**, *7*, 265–270. [CrossRef]
57. Cristescu, M.P.; Nerisanu, R.A.; Mara, D.A.; Oprea, S.V. Using market news sentiment analysis for stock market prediction. *Mathematics* **2022**, *10*, 4255. [CrossRef]
58. Rizinski, M.; Mishev, K.; Chitkushev, L.T.; Vodenska, I.; Trajanov, D. Using NLP transformer models to evaluate the relationship between ethical principles in finance and machine learning. In Proceedings of the 13th International Conference on Information Society and Technology (ICIST) Conference, Kopaonik, Serbia, 12–15 March 2023.
59. Atak, A. Exploring the sentiment in Borsa Istanbul with deep learning. *Borsa Istanb. Rev.* **2023**, *23*, S84–S95. [CrossRef]
60. Yang, H.; Li, K. Improving implicit sentiment learning via local sentiment aggregation. *arXiv* **2021**, arXiv:2110.08604.
61. Yang, H.; Zhang, C.; Liu, X.Y. PyABSA: A modularized framework for reproducible aspect-based sentiment analysis. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 5117–5122.
62. Kirange, D.; Deshmukh, R.R.; Kirange, M. Aspect-based sentiment analysis SemEval-2014 Task 4. *Asian J. Comput. Sci. Inf. Technol. (AJCSIT)* **2014**, *4*, 1.
63. Jiang, Q.; Chen, L.; Xu, R.; Ao, X.; Yang, M. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6280–6285.
64. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; de Clercq, O.; et al. SemEval-2016 Task 5: Aspect-based sentiment analysis. In Proceedings of the Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 19–30.
65. Mukherjee, R.; Shetty, S.; Chattopadhyay, S.; Maji, S.; Datta, S.; Goyal, P. Reproducibility, replicability and beyond: Assessing production readiness of aspect-based sentiment analysis in the wild. In *Advances in Information Retrieval, Proceedings of the 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 28 March–1 April 2021*; Proceedings, Part II 43; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 92–106.
66. Rajapaksha, S.; Ranathunga, S. Aspect detection in sportswear apparel reviews for opinion mining. In Proceedings of the 2022 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 27–29 July 2022; pp. 1–6.
67. Cooray, T.; Perera, G.; Kugathasan, A.; Alosius, J. Aspect-based sentiment analysis: Movie and television series reviews. In Proceedings of the International Workshop on Advanced Imaging Technology (IWAIT), Online, 5–6 January 2021; Volume 11766, pp. 615–620.
68. Boitel, E.; Mohasseb, A.; Haig, E. A comparative analysis of GPT-3 and BERT models for text-based emotion recognition: Performance, efficiency, and robustness. In *UK Workshop on Computational Intelligence*; Springer Nature: Cham, Switzerland, 2023; pp. 567–579.
69. Mughal, N.; Mujtaba, G.; Kumar, A.; Daudpota, S.M. Comparative analysis of deep neural networks and large language models for aspect-based sentiment analysis. *IEEE Access* **2024**, *12*, 60943–60959. [CrossRef]
70. Mahendru, S.; Pandit, T. SecureNet: A comparative study of DeBERTa and large language models for phishing detection. *arXiv* **2024**, arXiv:2406.06663.
71. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; Hu, X. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–32. [CrossRef]
72. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
73. Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. Zero-shot information extraction via chatting with ChatGPT. *arXiv* **2023**, arXiv:2302.10205.
74. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. Online Manuscript Released 20 August 2024. 2024. Available online: https://web.stanford.edu/~jurafsky/slp3 (accessed on 23 December 2024).
75. Chicco, D.; Jurman, G. The advantages of the matthews correlation coefficient (mcc), over f1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]

76. Rospocher, M.; Eksir, S. Assessing Fine-Grained Explicitness of Song Lyrics. *Information* **2023**, *14*, 159. [CrossRef]
77. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–38. [CrossRef]