Argument-Based Sentiment Analysis on Forward-Looking Statements

Chin-Yi Lin,[1] **Chung-Chi Chen,**[2] **Hen-Hsen Huang,**[3] **Hsin-Hsi Chen** [1]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] AIST, Japan
[3] Institute of Information Science, Academia Sinica, Taiwan
cylin@nlg.csie.ntu.edu.tw, c.c.chen@acm.org,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

This paper introduces a novel approach to analyzing the forward-looking statements in equity research reports by integrating argument mining with sentiment analysis. Recognizing the limitations of traditional models in capturing the nuances of future-oriented analysis, we propose a refined categorization of argument units into claims, premises, and scenarios, coupled with a unique sentiment analysis framework. Furthermore, we incorporate a temporal dimension to categorize the anticipated impact duration of market events. To facilitate this study, we present the Equity Argument Mining and Sentiment Analysis (Equity-AMSA) dataset. Our research investigates the extent to which detailed domain-specific annotations can be provided, the necessity of fine-grained human annotations in the era of large language models, and whether our proposed framework can improve performance in downstream tasks over traditional methods. Experimental results reveal the significance of manual annotations, especially for scenario identification and sentiment analysis. The study concludes that our annotation scheme and dataset contribute to a deeper understanding of forward-looking statements in equity research reports.

1 Introduction

The study of argument mining has been a focal point of comprehensive research (Toulmin, 2003; Cabrio and Villata, 2018; Lawrence and Reed, 2019), playing a crucial role in unraveling the complex views of authors or speakers. The foundational Toulmin argument model (Toulmin, 2003) classifies argument units into claims and premises (evidence) and dissects their interrelations. Nevertheless, such a categorization seems overly broad for interpreting the intricacies of forward-looking arguments, those primarily concerned with future analysis (Chen and Takamura, 2024). For example, when predicting future price trends, the Toulmin model treats past events and anticipated future scenarios as identical types of premises supporting an investor's claim. However, under the efficient-market hypothesis (Malkiel, 1989, 2003), events that have already transpired rapidly reflect in the market. Considering this, we propose a more detailed subdivision of premises into two categories: premises and scenarios. This approach aids in a deeper understanding of the author's future-oriented analysis.

Label	Example
Claim (Bearish)	We expect shares of Overweight-rated Apple to be under pressure in the near term.
Premise (Negative)	iPhone units were light, and the guidance for the Mar-Q implies continued softness, alongside higher OpEx.
Scenario (Collapse)	We think the iPhone air pockets reflect broader slowing in the smartphone market and company-specific factors.

Table 1: Example of argument-based sentiment labels.

Scenario planning is a critical process for conceptualizing and formulating comprehensive long-term plans (Amer et al., 2013). For instance, given the context of COVID-19 recovery, envisioning the future of public transit and shared mobility is a pressing issue for the transportation industry. To execute scenario planning, Shaheen and Wong (2021) engage numerous experts to deliberate on this topic. This underlines the fact that scenario planning still heavily relies on experts (Shaheen and Wong, 2021) and underscores the significance of identifying expert-written scenarios for understanding future projections. Given this factor, this paper introduces the task of scenario identification, along with the traditional task of argument unit identification, emphasizing the importance of scenarios authored by experts. As equity research reports typically delve into the future, we leverage them as our primary source.

We elevate this approach by merging argument

Findings of the Association for Computational Linguistics: ACL 2024, pages 13804–13815

mining with sentiment analysis, enabling a profound understanding of opinions and their interconnections. Deviating from traditional sentiment analysis that employs a generic label set (positive/negative), we propose distinct label sets for different argument units (claim, premise, and scenario). Table 1 shows an example in the proposed dataset. When making claims, investors reveal their personal perspectives (bullish/bearish/neutral). To support their claims, investors reference objective premises (positive/negative/neutral) and formulate scenario plans to contextualize these events. We utilize Dator's Four Generic Scenario Archetypes method (Dator and Dator, 2019) to categorize scenarios into four labels (continued growth/ steady state/ transformation/ collapse). This proposed argument-based sentiment analysis framework lays the foundation for future exploration of forward-looking arguments from a nuanced perspective.

Furthermore, when talking about the future, how long the impacts will continue becomes an important question (Tseng et al., 2023). For example, we may not use the earnings of a company in 2010 to assess the value of this company in 2024. To capture the duration of impacts implied in the arguments, we categorize the temporal dimension into: "within a month," "1-3 months," "4-6 months," "6-12 months," and "over a year". This classification is driven by the observation that most equity research reports discuss impacts spanning the upcoming year. By incorporating the temporal dimension, we distinguish between fleeting market sentiments, which are immediate reactions to news or short-term events, and more enduring financial analyses that reflect deeper market mechanisms and long-term company performance. Table 2 showcases examples by their anticipated impact durations. The first indicates a temporary rebound for Catcher in July, attributable to Sony Xperia casings, reflecting short-lived market fluctuations. In contrast, the second delves into the high-end smartphone market's anticipated saturation and a predicted 10% growth over 1-2 years, illustrating a sustained market outlook. Such demarcation provides clarity on the expected duration of an event's influence, enabling a deeper understanding of forthcoming market shifts.

In summary, this paper presents a novel annotation scheme for the automated interpretation of equity research reports by integrating the methodologies of argument mining and sentiment analysis. Addressing the lack of annotated professional analysis reports available to the research community, we introduce a comprehensive and expansive dataset named Equity Argument Mining and Sentiment Analysis (Equity-AMSA). This dataset covers a wide temporal range and volume, thus broadening the scope of our research community's endeavors. To conduct a thorough examination of the proposed Equity-AMSA, we aim to address the following three research questions (RQs):

Duration Label	Argument
1 to 3 months	Catcher may have a short-term technical rebound especially when July could be a strong month due to ramp for Sony Xperia casings.
Over a year	We think the high-end smartphone market is nearly saturated, and growth for the next 1-2 years will only be 10%. With intensifying competition from other tier-one players and entrance of lower-tier players, we think a long-term margin downtrend is inevitable for the smartphone market.

Table 2: Example for impact duration.

- (RQ1): To what extent can someone with domain knowledge provide annotations in this detailed manner?
- (RQ2): Why are fine-grained human annotations still necessary in an era where large language models demonstrate high performance in many NLP tasks?
- (RQ3): Can the proposed argument-based sentiment analysis scheme enhance performance in downstream tasks compared to traditional sentiment analysis and argument mining approaches?

2 Related Work

Financial news, earnings calls, and social media have been extensively used to create annotated datasets for model training. These annotations usually entail assigning sentiment labels, identifying entities, and performing argument analysis. For example, the Financial Phrase Bank dataset (Malo et al., 2014) comprises nearly 5,000 English sentences sourced from financial news. Each sentence is classified as positive, negative, or neutral, based on its emotional tone. StockTwits (Jaggi et al., 2021), a social media platform, offers labeled

tweets about companies, signaling sentiment as either bullish or bearish. This dataset encompasses various stocks and extends over several years of data collection. Alhamzeh et al. (2022) annotated argument units and structures in a dataset containing 804 documents from financial earnings calls. However, datasets specifically focusing on annotating argument-based equity research reports are seldom found. Despite these resources, none of the previous studies have amalgamated the concepts of argument mining, sentiment analysis, and scenario planning for a detailed understanding of financial opinions. This paper pioneers this research direction and introduces the first dataset for this purpose. Moreover, we turn our attention to equity research reports, a resource rarely included in previous open-access datasets, despite its pivotal role in the financial market. We anticipate that our dataset will prompt diverse discussions on opinion mining in financial documents.

3 Dataset

3.1 Data Source

We collected English equity research reports from Bloomberg Terminal.[1] Our data source comprises 1,876 analyst reports from the period between 2014 and 2022. According to the analysts' ratings, 47% of the reports are categorized as "Buy," 33% as "Neutral," and a mere 18% as "Sell." This distribution reveals a predisposition towards positive or neutral recommendations in analyst reports, potentially resulting from conflicts of interest, legal stipulations, or investor inclinations. The average word count for each report is 391 words, with the lengthiest report containing 735 words and the briefest one having 118 words. Typically, each report is structured into 3 to 4 paragraphs.

3.2 Annotation Task Guidelines

The primary objective of this annotation task is to identify the premise, scenario, and claim within the text. Annotators then assign various labels based on the type of statement, such as impact duration and sentiment. The labels are defined as follows:

Claim: This term refers to the forecasts or expectations posited by analysts regarding a company's growth and profitability. When labelling sentiments associated with these claims, annotators are instructed to discern whether the analyst's outlook is bullish, bearish, or neutral towards the company.

- Recognizing whether the analyst is bullish, bearish, or neutral towards the company
- Determining the duration of a bullish or bearish trend
- Deciding if the duration is explicitly stated in the report or requires subjective judgement

Premise: This pertains to events that have already transpired or are expected to transpire. In terms of sentiment labels on premises, annotators are requested to assess the event's influence on the company, whether it is positive, negative, or neutral.

- Differentiating whether the event is past or future-oriented
- Evaluating the event's positive, negative, or neutral impact on the company
- Indicating the impact duration from the publication date of the report
- Deciding if the impact duration is explicitly stated in the report or requires subjective judgement
- Assessing whether the event critically impacts or sparks a pivot in the company's future

Scenario: This encapsulates potential future events predicted by analysts. Concerning the sentiment labels on scenarios, annotators are tasked to gauge the implications for the company's future performance, i.e., whether it suggests continued growth, obstacles or decline, maintaining stability, or the prospect of significant changes or challenges. The labels are as follows:

- Determining the impact on the company's future performance: continued growth, facing obstacles or decline, maintaining stability, or encountering significant changes or challenges
- Identifying the scenario's impact duration from the report's publication date
- Deciding if the impact duration is explicitly stated in the report or requires subjective judgement
- Assessing whether the scenario critically impacts or sparks a pivot in the company's future

3.3 Inter-Annotator Agreement

Our annotation campaign involved five master students from the Department of Finance. We initially assigned a manageable set of reports to each annotator weekly, with the quality of their work assessed through inter-annotator agreement and weekly review meetings. In these meetings, we examined each instance that got inconsistent annotations, with the goal of establishing an acceptable

[1] https://www.bloomberg.com/professional/solution/bloomberg-terminal/

	Token-Based F1	Sentence-Based F1
Claim	0.601	0.765
Premise	0.792	0.796
Scenario	0.531	0.521

Table 3: Agreement of argument unit identification.

		Token-Based F1	Sentence-Based F1
Claim	Bullish	0.552	0.663
	Bearish	0.507	0.572
	Neutral	0.466	0.629
Premise	Positive	0.742	0.767
	Negative	0.653	0.616
	Neutral	0.366	0.462
Scenario	Continued Growth	0.500	0.398
	Steady State	0.239	0.289
	Collapse	0.297	0.351
	Transformation	0.148	0.224

Table 4: Agreement of argument-based sentiment.

level of uniformity among annotators. On average, each annotator annotated roughly 6,200 instances across 720 reports. Annotators are paid 30% higher than the legal minimum wage.

We gauge the inter-annotator agreement by computing the pairwise F1 Score between two annotators (Yang et al., 2018). Given that annotators are not confined to annotating at the sentence level and may annotate text spans arbitrarily, we estimate the F1 Score using the following two different definitions to evaluate the agreement:

Token-level F1 Score: In this definition, each token is treated as an annotation unit, implying that each token carries a label. We then determine the F1 agreement between the two annotators using the aforementioned method.

Sentence-level F1 Score: Here, we treat each sentence as an annotation unit. If an annotator splits a sentence into two annotations, for the purposes of calculating the F1 Score, both labels are regarded as potential gold annotations. That is, if there is an overlap between the annotations of the two annotators and the labels match, it is deemed correct. Subsequently, we compute the F1 agreement between the two annotators using the described method.

Table 3 presents the agreement level for the task of argument unit identification. The agreement for the scenario label is noticeably lower compared to the other labels. This discrepancy may arise due to analysts often referencing company revenue or profitability indicators while predicting future scenarios, which could potentially be confused with the claim label. Table 4 depicts the level of agreement for the task of argument-based sentiment anal-

	Token-Based F1	Sentence-Based F1
< 1 month	0.240	0.270
1 to 3 month	0.588	0.520
4 to 6 month	0.517	0.472
7 to 12 month	0.595	0.596
> 1 year	0.628	0.654

Table 5: Agreement of argument-based impact duration.

	Kappa score	Percentage
Explicit Annotations	0.613	65%
Implicit Annotations	0.302	35%

Table 6: Agreement on Explicit and Implicit Annotations of Impact Duration.

ysis. We have observed a lower agreement in the sentiment labeling of premises and claims, particularly concerning the assignment of the neutral label. This variation can be attributed to differing sensitivity levels among annotators when determining whether an event is positive/negative or neutral, and to the presence of both neutral and positive/negative narratives within a single annotation, leading to divergent views among annotators. In the sentiment labeling of scenarios, we notice a significantly lower agreement level than for premises and claims. This is primarily due to inconsistencies in identifying text spans that should be labeled as scenario labels within the same report, leading to an indirect impact on the agreement level for sentiment labels of scenarios.

Table 5 illustrates the agreement on argument-based impact duration. For short-term impacts, like "within a month," the agreement is relatively low, suggesting such impact duration might be ambiguously mentioned in reports. In contrast, broader durations such as "over a year" yield higher agreement. This suggests that more extensive time frames are frequently described with greater clarity and emphasis, facilitating easier identification by annotators. In light of the observation, we introduce "explicit" and "implicit" labels to allow annotators to indicate whether the impact duration could be clearly derived from the text or required subjective judgment. In Table 6, we also calculate the agreement based on the two labels independently. For instances where both annotators categorized as "implicit," the Cohen's Kappa coefficient (Cohen, 1960) value is 0.302, indicating fair agreement. However, in situations where both annotators identified as "explicit," the Kappa value increases substantially to 0.613, signifying substantial agreement. In total, we collected 65% of ex-

	Cohen's Kappa
Argument Unit	0.713
Claim Sentiment	0.774
Premise Sentiment	0.809
Scenario Sentiment	0.682
Impact Duration	0.417

Table 7: Agreement on argument units, sentiment, and impact duration using Cohen's Kappa.

	Sentiment Label	Training	Development	Test
Claim	Bullish	3,831	426	439
	Bearish	2,397	267	320
	Neutral	1,348	150	170
Premise	Positive	5,058	562	1,965
	Negative	4,120	458	1,387
	Neutral	1,456	162	149
Scenario	Continued Growth	2,431	270	629
	Steady State	504	56	110
	Collapse	1,927	214	417
	Transformation	453	50	52

Table 8: Statistics of the experimental dataset — Argument & Sentiment.

Duration Label	Training	Development	Test
<1 month	655	73	41
1-3 month	5,258	584	1,044
4-6 month	3,589	399	457
7-12 month	9,004	1,000	2,121
>1 year	5,511	612	1,308

Table 9: Statistics of the experimental dataset — Impact Duration.

plicit annotations and 35% of implicit annotations. To enhance consistency and address the aforementioned issues, we also introduced some measures. We instructed the annotators to carefully review the text for any time-related terms and to specify precise time intervals, thereby minimizing ambiguities related to time spans. Additionally, we've noticed that some inconsistencies between annotators are due to instances containing two distinct labels. As a result, we instruct annotators to break down the sentence into its smallest unit that explicitly represents a single label to minimize disagreement. Specifically, when annotators encounter a sentence containing two distinct labels, such as positive and negative sentiment labels in the same event, they are required to divide the sentence into two individual instances.

In addition to the F1 Score, we also utilize the Cohen's Kappa coefficient specifically to measure the overall agreement on instances that were labeled by both. The results of Cohen's Kappa are presented in Table 7. These results provide the answer to RQ1. The annotation results indicate that the proposed tasks are inherently subjective, particularly in terms of scenario identification and scenario sentiment analysis. This observation aligns with the characteristics of the financial market, where investors may interpret identical research reports differently, influencing their market decisions (buy/sell). During review meetings, reaching consensus on specific cases was challenging due to varying interpretations among annotators. For instance, the statement: "Dow reduced its capital expenditures from $2 billion in 2019 to approximately $1.2 billion in 2020." was interpreted positively by Annotator A, highlighting the potential for short-term cost savings and profitability enhancements from reduced capital expenditures. Conversely, Annotator B viewed it negatively, suggesting that such a reduction in investment could hinder growth prospects and competitiveness. This discrepancy exemplifies the challenge in determining the "correct" perspective, as both viewpoints are valid. Recent research underscores the value of learning from multiple perspectives, a promising direction that diverges from the traditional supervised learning paradigm focused on a single ground truth (Bender and Friedman, 2018; Basile et al., 2021). Consistent with this approach, our dataset includes annotations from multiple annotators, rather than providing definitive labels. To support research aimed at a singular ground truth, we organize weekly meetings with annotators to discuss cases with divergent annotations and assign final labels based on these discussions.

3.4 Dataset Statistics and Analysis

We have annotated a total of 37,416 instances, mostly using sentences as the annotating units, classifying these instances into premises, scenarios, or claims. Moreover, we have recognized 10,485 groups that correspond to the count of groups in which premises, scenarios, and claims are deemed correlated (i.e., group(premise, scenario, claim)). Thus, each group, anchored by a claim, represents a comprehensive argument, with supportive premises and potential scenarios related to the central claim. Tables 8 and 9 provide a comprehensive statistical breakdown for each label type. Additionally, we have enumerated the top 5 most frequently occurring indicators annotated within claims. These indicators— "Price target," "EPS (earnings per share)," "revenue," "YoY" (Year on Year), and "GM" (Gross Margin)—are critical metrics often scrutinized by

	Argument Unit Identification				Impact Duration Inference			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
ChatGPT(Zero-Shot)	0.723	0.739	0.723	0.725	0.543	0.631	0.557	0.532
ChatGPT(Few-Shot)	0.754	0.830	0.766	0.779	0.565	0.582	0.565	0.570
GPT-4(Few-Shot)	0.774	0.824	0.819	0.812	0.573	0.667	0.573	0.611
BERT	0.902	0.901	0.903	0.902	0.757	0.757	0.757	0.756
FinBERT	0.899	0.901	0.898	0.901	0.763	0.765	0.763	0.763
RoBERTa	**0.905**	**0.907**	**0.905**	**0.906**	**0.780**	**0.778**	**0.780**	**0.777**

Table 10: Experimental results of argument unit identification and impact duration inference.

	Claim				**Premise**				**Scenario**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
ChatGPT(Zero-shot)	0.819	0.862	0.825	0.833	0.875	**0.944**	0.882	0.901	0.774	0.891	0.773	0.817
ChatGPT(Few-Shot)	0.883	0.886	0.883	0.884	0.883	0.941	0.883	0.906	0.840	**0.915**	0.840	0.870
GPT-4(Few-Shot)	0.919	0.923	0.919	0.920	0.923	0.918	0.923	0.912	0.755	0.882	0.755	0.809
BERT	0.927	0.932	0.929	0.930	0.912	0.901	0.918	0.911	0.866	0.883	0.870	0.871
FinBERT	**0.929**	0.930	0.930	0.930	0.904	0.906	0.901	0.903	0.872	0.861	0.875	0.862
RoBERTa	0.922	**0.933**	**0.932**	**0.931**	**0.925**	0.919	**0.925**	**0.920**	**0.884**	0.904	**0.885**	**0.893**

Table 11: Experimental results of argument-based sentiment analysis.

analysts when evaluating companies and formulating recommendations.

4 Experiment

4.1 Baseline Models and Results

For our experiment, we utilized pre-trained models such as BERT (Devlin et al., 2019), FinBERT (Araci, 2019), RoBERTa (Liu et al., 2019), GPT-4[2] and ChatGPT (GPT-3.5)[3] as our baselines. Specifically, we fine-tune the models, namely BERT, FinBERT, and RoBERTa, on our tasks, with an added linear layer as a classifier leveraging the final hidden state associated with the [CLS] token. The inputs for sentiment analysis and argument unit identification are instances (a text span from a report) from our dataset. For impact duration inference, the inputs are instances (a text span from a report) along with the report's publication date. The corresponding ground truth labels are used as outputs during model training. It is important to note that, in this paper, the test set we utilized for evaluation comprises only instances that have achieved full agreement among the annotators. The models were fine-tuned using a learning rate of 2e-5, weight decay of 0.01, batch size of 32, 5 epochs, and the AdamW optimizer. To assess the performance of these refined models, we utilized accuracy, precision, recall, and F1-score as evaluation metrics.

Table 10 presents the results on the argument unit identification task. RoBERTa exhibits superior performance among all models and achieves an accuracy and F1-score above 90%. Given that ChatGPT and GPT-4 operate under zero-shot and few-shot settings, and we provide guidelines, including some demonstrations for each label in the few-shot setting, to request them to generate predictions, our intention in showcasing the performance of ChatGPT and GPT-4 is not for comparison with supervised models. The performance detailed in Table 10 suggests that GPT-4 can achieve substantial agreement with human annotators. That is, when the annotators' agreement level reaches 100%, GPT-4 attains 81.2% by following the same guidelines.

Table 10 also presents the results from the argument-based impact duration inference task. RoBERTa consistently outperforms other models across all metrics. With an accuracy, precision, recall, and F1 Score of approximately 78%. Additionally, we carry out an ablation study to understand the importance of including the report's publication date when fine-tuning the models. The results reveal that integrating the publication date lead to an approximate 0.1 increase in the F1 Score. This underscores the significance of the report's publication date in assessing the validity period of analysts' forecasts and the duration of event impacts.

Further analysis of ChatGPT's performance on specific labels indicates a pattern of low precision combined with high recall. This trend is especially pronounced for the labels: "Neutral" of premise (precision: 0.25, recall: 0.78), "Neutral" of claim (precision: 0.54, recall: 0.86), and "Steady State" of scenario (precision: 0.19, recall: 0.73). In essence, while ChatGPT often identifies a broad range of potential neutral arguments (high recall),

[2] https://openai.com/index/gpt-4-research/
[3] https://openai.com/blog/chatgpt

	Training Set	Argument Unit Identification				Impact Duration Inference			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
RoBERTa	Human	**0.905**	**0.907**	**0.905**	**0.906**	**0.780**	**0.778**	**0.780**	**0.777**
BERT	ChatGPT	0.651	0.729	0.652	0.655	0.451	0.463	0.462	0.420
FinBERT		0.657	0.735	0.658	0.661	0.452	0.458	0.459	0.433
RoBERTa		0.686	0.745	0.686	0.689	0.466	0.466	0.471	0.441

Table 12: Experimental results of argument unit identification and impact duration inference using ChatGPT labels for training.

	Claim				**Premise**				**Scenario**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
RoBERTa (Human)	**0.922**	**0.933**	**0.932**	**0.931**	**0.925**	0.919	**0.925**	**0.920**	**0.884**	**0.904**	**0.885**	**0.893**
BERT (ChatGPT)	0.769	0.860	0.770	0.790	0.897	0.920	0.900	0.910	0.781	0.830	0.780	0.790
FinBERT (ChatGPT)	0.777	0.855	0.777	0.793	0.888	0.920	0.889	0.901	0.803	0.852	0.804	0.821
RoBERTa (ChatGPT)	0.801	0.856	0.802	0.811	0.909	**0.931**	0.910	0.919	0.821	0.847	0.822	0.828

Table 13: Experimental results of sentiment analysis using ChatGPT labels for training.

it is less accurate in ensuring that these identifications are correct (low precision). This suggests that ChatGPT tends to misclassify polarized arguments as neutral, resulting in numerous false positives.

Table 11 displays the experimental results of the argument-based sentiment analysis tasks. Overall, RoBERTa continues to outperform the other models when evaluated based on the F1-score. In the argument-based sentiment analysis task, the performance gap between supervised models and ChatGPT narrows, particularly in the sentiment analysis of premises. Our results underscore the potential of employing ChatGPT for sentiment analysis tasks. Nevertheless, regardless of the tasks, a performance gap persists between supervised models and ChatGPT, which affirms the necessity of the proposed manually annotated dataset.

4.2 Significance of Human Annotation

The recent surge in discussions has revolved around the potential of LLMs as substitutes for human annotators in both dataset construction (Latif et al., 2023) and evaluation (Chiang and Lee, 2023). To answer RQ2, we delve into the extent to which we can depend on LLM-generated labels for supervised model training. By contrasting these models with those trained using human annotations, we aim to gauge the efficacy achieved using our proposed dataset. To this end, we construct a new training set, which annotations are provided by ChatGPT. The test set is the same as that in Section 4.1. We follow the same experimental setting as shown in Section 4.1 to train BERT, FinBERT, and RoBERTa with the new training set with ChatGPT-generated labels. It is worth noting that ChatGPT uses the same guidelines as human annotators, and the instances and size in the new training set are the same as the human-annotated training set.

Table 12 presents the performance of models on the argument unit identification task. The outcomes suggest that using manually-annotated labels consistently yields superior performance, irrespective of the underlying language model. When contrasted with Zero-shot ChatGPT performance, it becomes evident that auto-generating additional instances for training supervised models doesn't necessarily enhance the task performance. Turning our attention to Table 12, the trends reiterate the significance of manual annotations encapsulated in our Equity-AMSA.

Nevertheless, the insights drawn from the argument-based sentiment analysis task differ to some extent. Table 13 delineates the performance metrics for the argument-based sentiment analysis task. Results for both claim and scenario sentiment analyses emphasize the pivotal role of manual annotation. Notably, there's a marked performance disparity regardless of the model in play. In contrast, premise sentiment analysis exhibits distinct trends. Models trained on ChatGPT-generated labels achieve F1 Scores comparable to those fine-tuned with manually-generated labels. Furthermore, supervised models leveraging ChatGPT-generated labels outperform the zero-shot ChatGPT. This suggests the latent potential of ChatGPT-derived labels for premise sentiment analysis.

Synthesizing the insights from this section, it's evident that manual annotations predominantly elevate performance across tasks in our proposed dataset, with the lone exception being premise sentiment analysis. Given that a majority of extant research (Devitt and Ahmad, 2007; Hamborg

	Model	Feature	Accuracy	Precision	Recall	F1 Score
Zero-Shot	PaLM 2	-	0.608	0.609	0.608	0.591
	Gemini Pro	-	0.614	0.621	0.614	0.615
	ChatGPT	-	0.635	0.644	0.635	0.635
	GPT-4	-	0.640	0.662	0.640	0.638
Supervised	Longformer	-	0.705	0.723	0.705	0.699
	Llama-2-7B	-	0.696	0.708	0.696	0.696
	Mistral-7B	-	0.725	0.724	0.725	0.725
	GNN	-	0.731	0.732	0.731	0.728
		Conventional Argument Units	0.737	0.747	0.737	0.737
		Proposed Argument Units	0.749	0.761	0.749	0.741
		VADER Sentiment	0.738	0.739	0.738	0.738
		Argument-Based Sentiment	0.775	0.776	0.775	0.773
		Impact Duration	0.757	0.756	0.757	0.757
		All Labels in Equity-AMSA	**0.798**	**0.801**	**0.798**	**0.796**

Table 14: Results of predictability assessment.

and Donnay, 2021) primarily orbits around sentiment tasks akin to premise sentiment analysis in this study, our findings arguably introduce a nuanced, fine-grained sentiment analysis task centered around forward-looking statements, namely claims and scenarios.

5 Predictability Assessment

5.1 Experimental Setup

In equity research reports, professional analysts forecast a company's future performance, and the realization of these forecasts is a criterion for assessing their analysis (Zong et al., 2020). Inspired by Chen et al. (2019), each report encapsulates its analysis into a price target, which is the anticipated stock price level the analysts predict the company will achieve. Given the quarterly release of financial reports by companies, analysts update their forecasts to incorporate significant changes. We aim to evaluate the predictability of equity research reports by identifying the realization of the stated price targets within a three-month period. We used 1,775 reports for this experiment. For each report, we extracted historical stock closing prices from Yahoo Finance and assessed whether the price target was met within the subsequent three months. The dataset was split into a training set (80%) and a test set (20%). We prompted LLMs (ChatGPT, GPT-4, PaLM 2[4], and Gemini Pro[5]), and employed language model-based methods, Longformer (Beltagy et al., 2020), and graph neural network (GNN)-based method, SAGEConv architecture (Hamilton et al., 2017) as our baselines.

The aim of using Longformer is to address token size limitations. For fine-tuning Llama-2-7B and Mistral-7B models, we employed the parameter-efficient fine-tuning method LoRA, configuring the settings with a *lora_rank* of 256 and a *lora_alpha* of 512. Regarding GNN, we created an opinion graph for each report to predict price target realization. We generated 1,775 graphs, with an average of 10.55 nodes and 13.53 edges per graph. The graph is constructed based on the "group" mentioned in Section 3.4. We connect all correlated premises and scenarios to the corresponding claim, noting that these premises or scenarios are not constrained to only one claim; they can be connected to more than one claim if applicable. Additionally, all claims are connected to a virtual node to facilitate information exchange between nodes. The SAGEConv architecture (Hamilton et al., 2017) was employed to learn node representations by aggregating information from neighboring nodes. Initially, we encode each argument unit (e.g., premises, scenarios, or claims), combined with additional data such as impact duration or sentiment labels, to form graph nodes. We then applied two layers of SAGEConv for node refinement. A comprehensive graph representation was obtained by averaging all node representations, which was then used as input for a linear classifier.

5.2 Experimental Results

Table 14 presents the experimental results. The distribution of achieved price targets or not is 45.61% (achieved) and 54.39% (not achieved). Although LLMs can somehow identify the predictability of the given report, they still perform worse than supervised models. The basic GNN outperforms the Longformer, Llama-2-7B, and Mistral-7B. Enhancements to the GNN architecture with the inclu-

[4] https://ai.google/discover/palm2/

[5] https://deepmind.google/technologies/gemini/pro/

	Precision	Recall	F1
Achieved	0.822	0.712	0.763
Not Achieved	0.783	0.871	0.824

Table 15: Label-based Evaluation.

sion of all proposed features significantly boost its performance. To address RQ3, we contrast our annotation schemes against the sentiment labels (positive, negative, neutral) derived from the VADER toolkit (Hutto and Gilbert, 2014) and a GNN utilizing only conventional argument unit definitions (claim/premise) for comparison with our extended schemes (argument-based sentiment analysis and claim/premise/scenario). The findings indicate that our annotation schemes enhance performance in evaluating equity research reports, surpassing previous approaches in sentiment analysis and argument mining. Table 15 shows the label-based evaluation of the GNN model using all labels in Equity-AMSA, and it indicates that the model has high precision in identifying the reports that have high predictability and high recall in filtering out the reports that have low predictability.

5.3 Human Performance

To compare the ability of predictability assessment between models and human beings, we conducted a small-scale experiment with three master students from the Department of Finance who are experienced in financial analysis and have academic backgrounds in understanding financial text. We utilize a total of 75 reports, and evaluate the performance of Human, GPT-4 (zero-shot), Mistral-7B, and GNN using all labels in Equity-AMSA in predicting whether a target price would be achieved in the following three months, based on the information in the reports. Table 16 indicates that GPT-4, Mistral, and GNN outperform humans in this task. It shows the potential and capacity of the models to identify reliable reports and filter out unconvincing ones from a large volume of reports, and it helps investors quickly obtain accurate insights, viewpoints, and recommendations on the performance and potential of companies or financial assets.

Moreover, we evaluated human performance on problem sets of varying difficulty, where *difficulty* refers to how challenging the problems are for models. For example, Easy problems are those reports correctly predicted by all three models (GPT-4, Mistral, and GNN using all labels in Equity-AMSA), while Hard problems are those reports

	# correct predictions
Human	40
GPT-4 (zero-shot)	43
Mistral-7B	47
GNN w/ all labels	47

Table 16: Human evaluation of predictability assessment.

	# correct predictions by Human
Easy	16 out of 25
Medium	9 out of 25
Hard	15 out of 25

Table 17: Human evaluation of predictability assessment based on varying levels of difficulty.

that all three models incorrectly predicted. Each category contained 25 questions. As shown in Table 17, humans correctly predicted the outcomes for 16 out of the 25 reports in the easy problems, and for 15 out of the 25 reports in the hard problems. This indicates that humans excel in some cases where models struggle, suggesting that there are unique strengths individually in predictive models and humans for analyzing analysts' opinions and making predictions, which could result from their different available information, background knowledge or capabilities. This implies the potential to achieve complementary team performance with Human-AI collaboration.

6 Conclusion

This paper proposes a novel and comprehensive approach to the automated interpretation of equity research reports by integrating argument mining with sentiment analysis, and introduces the Equity-AMSA dataset. Our results show the indispensable role of human annotation in maintaining high-quality data for training machine learning models. Despite the advancements in LLMs, our findings reveal that manual annotations outperform LLM-generated labels in most tasks, highlighting the nuanced understanding humans bring to the interpretation of financial texts. Moreover, the predictability assessment through experimental evaluations demonstrates the superior performance of our proposed annotation schemes and the importance of task-specific features in enhancing the analysis of equity research reports. Our work serves as a foundation for future exploration, encouraging further investigation into the integration of argument mining and sentiment analysis for the nuanced interpretation.

Limitations

A primary limitation of this paper lies in its singular focus on one type of financial document. As elucidated by Chen et al. (2021), financial sources can be categorized into four clusters based on their providers: company/managers, professional investors, social media users, and journalists. Our choice to center on professional investors' reports is driven by several considerations. Managers often provide limited insights into a company's prospective operations, while journalists typically prioritize factual reporting over forecasting. This constraint curtails our ability to delve into forward-looking statements from these two sources. Social media data, on the other hand, is often cluttered and informal, making it less suitable for immediate study. Thus, we earmark it for potential future research. Given these constraints, we posit that initiating our analysis with professional reports offers a reasonable starting point for the nuanced, argument-based sentiment analysis tasks we propose. A secondary limitation pertains to the linguistic scope of our dataset, which exclusively features English. This may curtail discussions encompassing other languages. However, since many models demonstrate peak performance in English, we deem it a pragmatic starting point. Future research endeavors can emulate our task framework and methodologies to delve into other languages, juxtaposing their findings with the insights from this study.

Ethical Considerations

Publishing content on platforms distinct from the original site requires permissions or copyright transfer, hence, researchers often release URLs or tweet IDs with their annotations. Adhering to this, our Equity-AMSA dataset releases annotations with equity report file names under the CCBY-NC-SA 4.0 license. We also offer codes to reconstitute the dataset from original reports.[6] Researchers can independently download reports from the Bloomberg Terminal and apply our codes to rebuild the dataset.

Acknowledgement

This work was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-113L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

[6] https://github.com/CYXup6/Equity_AMSA

References

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Amer, Tugrul U Daim, and Antonie Jetter. 2013. A review of scenario planning. *Futures*, 46:23–40.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From opinion mining to financial argument mining*. Springer Nature.

Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference*, pages 3498–3502.

Chung-Chi Chen and Hiroya Takamura. 2024. Term-driven forward-looking claim synthesis in earnings calls. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15752–15760, Torino, Italia. ELRA and ICCL.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Jim Dator and Jim Dator. 2019. Alternative futures at the manoa school. *Jim Dator: A Noticer in Time: Selected work, 1967-2018*, pages 37–54.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Clayton Hutto and Eric Gilbert. 2014. 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, pages 216–225.

Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, and Matloob Khushi. 2021. Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, 4(1).

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint arXiv:2307.06090*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Burton G Malkiel. 1989. Efficient market hypothesis. *Finance*, pages 127–134.

Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Susan Shaheen and Stephen Wong. 2021. Future of public transit and shared mobility: Scenario planning for covid-19 recovery.

Stephen E Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A lightweight collaborative text span annotation tool. In *Proceedings of ACL 2018, System Demonstrations*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Shi Zong, Alan Ritter, and Eduard Hovy. 2020. Measuring forecasting skill from text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5317–5331, Online. Association for Computational Linguistics.

A Prompt for LLMs

We used the prompt below for the experiments of PaLM 2, Gemini Pro, ChatGPT, and GPT-4 in Table 14.

Based on the following equity research report, predict the likelihood of the stated target price being achieved within the next three months, carefully considering the opinions presented in the report. Your response should be either "The target price is likely to be realized" or "The target price is unlikely to be realized." Equity Research Report: <report>

We used the prompts below for the experiments of ChatGPT in Table 10.

You are a financial analyst. I will give you a sentence from an equity research report. Please classify the sentence into premise, scenario or claim:

Premise: Events that have occurred or are expected to occur Scenario: Possible future events envisioned by analysts Claim: Analysts' expectations or forecasts regarding company growth and profitability Sentence: <sentence> The sentence can be classified as

You are a financial analyst. I will give you a sentence from an equity research report. Please evaluate the impact duration of the statement <sentence>, given that the report was published on <publish date>. Question: How long will the impact of the statement last? (A) within a month (B) 1-3 month (C) 4-6 month (D) 7-12 month (E) Over a year

We used the prompts below for the experiments of ChatGPT in Table 11.

You are a financial analyst. I will give you a sentence from an equity research report. Please classify the sentence into "Bullish", "Bearish" or "Neutral". Sentence: <sentence> The sentence can be classified as

You are a financial analyst. I will give you a sentence from an equity research report. Please classify the sentence into "Positive", "Negative" or "Neutral". Sentence: <sentence> The sentiment of the sentence is

You are a financial analyst. I will give you a sentence from an equity research report. Please classify the sentence into "Steady State", "Continued Growth", "Collapse" or "Transformation". The definitions of the four types are as follow: Steady State: maintaining stability Continued Growth: continued growth Collapse: encountering obstacles or decline Transformation: facing significant changes or challenges Sentence: <sentence> The sentence can be classified as