Enhancing Investment Opinion Ranking through Argument-Based Sentiment Analysis

Chung-Chi Chen,[1] **Hen-Hsen Huang,**[2] **Hsin-Hsi Chen** [3]
Hiroya Takamura,[1] **Ichiro Kobayashi,**[4] **Yusuke Miyao**[5]
[1] AIST, Japan [2] Institute of Information Science, Academia Sinica, Taiwan
[3] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[4] Ochanomizu University, Japan [5] The University of Tokyo, Japan
c.c.chen@acm.org, hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw,
takamura.hiroya@aist.go.jp,koba@is.ocha.ac.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

In the era of rapid Internet and social media development, individuals readily share their investment opinions online. The overwhelming volume of such opinions makes comprehensive evaluation impractical, highlighting the need for an effective recommendation system that can identify valuable insights. To address this challenge, we propose an argument-based sentiment analysis framework that incorporates a new perspective on opinion strength. Our approach introduces the concept of a Fuzzy Strength Degree (FSD), derived from the difference between analysts' target and closing prices, to quantify the intensity of opinions. By integrating argument mining techniques, we further decompose each opinion into claims and premises, examine their relationships, and use these structures to evaluate the persuasive strength of the arguments. This dual strategy allows us to rank both professional and amateur investor opinions without relying on user history or social signals. Experiments show that our method works best for analyst reports, while on social media, simpler approaches based on wording and professionalism features perform better. Moreover, our analysis of professional analysts' and traders' behaviors reveals that top-ranked opinions are more likely to influence subsequent market actions. These findings demonstrate that argument structure and quantified opinion strength provide a novel and reliable foundation for investment opinion recommendation.

1 Introduction

In recent years, online platforms catering to diverse interests have proliferated, ranging from personal updates on X (formerly Twitter) and Instagram to professional discourse on MathOverflow[1] and Stack Overflow[2]. Specialized platforms also exist for sharing perspectives on specific topics, such as iDebate[3] for user debates and various e-commerce sites for product reviews. Among these, platforms focusing on future-oriented user predictions, especially in the investment domain, play a unique role. Sites like Yahoo Finance[4] and Reddit's WallStreetBets[5] are popular for sharing professional and amateur investment opinions. However, as Xiao (2015) notes, only about 20% of investors see profits from stock market activities, indicating the challenge of discerning valuable investment opinions amidst a deluge of information.

Previous research has primarily focused on predicting stock price movements using publicly available opinions (Sawhney et al., 2020; Xu and Cohen, 2018; Bollen and Mao, 2011). However, few have concentrated on extracting the subset of opinions that are likely to be profitable. Current methods rely on analyzing either the characteristics of the authors or the content of their posts. Author-centric analysis, despite its usefulness in identifying potential sources of profitable insights, faces several challenges, such as the lack of historical data for new users, the anonymity and account variability on social media, and the inconsistency of user information across platforms. To address these challenges, this paper proposes a text-centric approach to analyze opinions and identify those likely to result in profitable investments. This involves using historical data to train models for predicting profitable outcomes and extracting linguistic features to sort out useful opinions. We introduce a novel method that leverages the concept of argument mining to analyze opinions. This approach decomposes opinions into premises and claims, examining their interrelations to provide a more nuanced understanding of the opinion expressed.

Argument mining, as described by Mochales and Moens (Mochales and Moens, 2011), is the pro-

[1] https://mathoverflow.net/

[2] https://stackoverflow.com/

[3] https://idebate.org/

[4] https://finance.yahoo.com/research/

[5] https://www.reddit.com/r/wallstreetbets/

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1305–1315
December 20-24, 2025 ©2025 Association for Computational Linguistics

cess of extracting and identifying the argumentative structure of a given text, particularly in subjective and persuasive narratives. Investment opinions, which forecast market futures, inherently contain subjective viewpoints of investors. To persuade readers, it is essential to support claims with evidence (premises). This paper goes beyond merely distinguishing between premises and claims in a narrative; we introduce the concept of measuring the strength of an argument. A novel aspect of our methodology is the use of price targets set by professional analysts as proxies for argument strength. These targets, representing the analysts' future price estimations, enable the quantification of the argument's strength. Additionally, this study investigates the influence of objective evidence and subjective opinion in forming investment recommendations. Through argument mining, we distinguish between factual statements and personal views within an opinion, thereby enhancing the analysis. Our approach, termed argument-based opinion analysis, presents a novel perspective in the domain of investment opinion recommendation.

Our research also extends into the realm of professional behavior, examining changes in analysts' views and trading patterns among professional traders. We find that opinions ranked highly by our method are more correlated to analysts' revisions and traders' decisions. This aspect of professional behavior, though crucial, has been largely overlooked in previous studies.

In summary, the contributions of this work are threefold: (1) Introduction of a novel approach for assessing strength degree of opinion using price targets set by professional analysts. (2) Development and evaluation of various strategies based on argument mining notions for investment opinion recommendation. (3) Comprehensive analysis of the implications of differently ranked opinions on profitability, risk, and professional trading behaviors.

2 Related Work

Investment decision-making based on online opinions has been a topic of long-standing discussion. The majority of studies have harnessed the 'wisdom of the crowd' by aggregating available opinions (Sawhney et al., 2020; Xu and Cohen, 2018; Bollen and Mao, 2011), with few exploring the utility and relevance of individual opinions. In this paper, we take the latter's perspective, proposing a novel direction for the assessment of individual opinions.

Historically, research has examined this issue from two angles: author-based and content-based. Bar-Haim et al. (2011) focused on identifying experts based on their past performance. Tu et al. (2018) prioritized authors' social popularity to sift through investors' opinions. However, these methods encounter a significant cold-start problem where new users lack historical data and users can manipulate their online presence through account changes or deletion of past posts. Moreover, it can be challenging to obtain the necessary author-specific data in various scenarios. For instance, the social popularity of professional analysts is not readily available for assessing their opinions. As a result, we propose a content-centric approach that can be universally applicable.

Related studies have also examined opinions based on content. Zong et al. (2020) used BERT (Devlin et al., 2019) to classify the accuracy of analyst reports according to company earnings forecasts and actual earnings. This was treated as a binary classification task, wherein a given report was either in the top or bottom K groups, selected from 16,044 reports. However, we argue that such a coarse setting is unsuitable for recommending a practical number of reports for investors. Furthermore, social media platforms typically lack historical performance records, making it challenging to compile sizeable training data. In contrast, our method can rank opinions on a continuous scale (from 0 to 1) using a limited number of reports with automatically annotated information.

Chen et al. (2021) put forth an approach to identify high-predictive-power opinions on social media platforms, wherein the predictive power of posts was determined by the similarity of their sentences to those written by professional analysts. However, this approach faces a significant issue in that it cannot rank professional investors' opinions, as all sentences in analyst reports are authored by professionals. We offer a novel method to overcome this limitation.

Following the idea of previous work, Chen et al. (2024b) propose a professionalism-aware *pre-finetuning* scheme placed between pre-training and task fine-tuning to improve ranking of investor opinions by profitability. Concretely, they introduce two auxiliary tasks that teach language models to distinguish professional vs. amateur writing: a *sentence-level* discriminator (was a sentence written by a professional analyst?) and a *word-level* discriminator (is a token professional-leaning). The key advantage over prior "expert-like sentence counting" approaches is coverage: their models assign scores to *all* opinions, including posts without explicit expert-like sentences. They also compare *pairwise* ranking against regression and report that pairwise training yields better overall ranking quality (e.g., higher DCG/nDCG, stronger top-decile MPP) on analyst reports and social media benchmarks. (Chen et al., 2024b) This makes their framework a natural and competitive baseline for our setting.

3 Task Formulation and Dataset

This section defines the recommendation task and describes the datasets used to evaluate our proposed methods. We first formalize the investment opinion recommendation problem, followed by the introduction of the datasets collected from professional and social media sources.

3.1 Investor Opinion Recommendation

This paper aims to identify potentially profitable investment opinions from a vast array of perspectives. Rather than merely sifting through the plethora of opinions, our objective is to pinpoint and recommend those with the highest potential for yielding profitable outcomes to investors. Considering that most investor opinions only provide information on initiating a trading position, such as whether to go long or short on a stock, but don't suggest when to close the position, assessing the actual return achievable by following the opinion on the market becomes challenging. As such, we adopt the approach of previous work (Chen et al., 2021), using the Maximum Possible Profit (MPP) as the evaluation metric for potential return. This approach circumvents the problem of determining when to close the position and provides a reasonable estimation of the potential return from an opinion. Below are the equations used to calculate the MPP of both bullish and bearish opinions for a stock on day t:

$$MPP_{bullish} = \max_{t+1 \leq i \leq t+T} \frac{H_i - O_{t+1}}{O_{t+1}} \tag{1}$$

$$MPP_{bearish} = \min_{t+1 \leq i \leq t+T} \frac{O_{t+1} - L_i}{O_{t+1}} \tag{2}$$

In the above equations, O_t represents the opening price on day t, with H_t and L_t denoting the highest and lowest prices on day t respectively. Following Chen et al. (2021), we set T as 60 trading days, or roughly three months. Ultimately, our goal is to devise a method to identify opinions that could lead to a higher MPP.

3.2 Investor Opinion Sets

Contrary to prior studies that focused solely on either formal reports (Zong et al., 2020) or social media posts (Chen et al., 2021), our work proposes a method capable of handling both. Consequently, our experiments involve two sets of opinions: analyst reports and social media posts. This section details these opinion sets. It should be noted that these sets are used exclusively for evaluating recommendation approaches and not during any training steps.

We assembled a collection of 2,280 professional analysis reports from the Bloomberg Terminal. Written in Chinese, these reports delve into stocks listed on the Taiwan stock market, covering a total of 513 different stocks. The release date for each report, indicated in the file name, enables us to easily ascertain t. Once t is known, we align it with the price data based on date information, allowing us to calculate the MPP for each report. This collection is henceforth referred to as the PAR set.

A set of social media posts was obtained from Chen et al. (2021). Written in Chinese, these posts were sourced from PTT, a well-known social media platform in Taiwan.[6] Chen et al. (2021) manually annotated the MPP of each post, considering the sentiment (bullish/bearish) expressed in these posts and the stocks discussed. This collection is referred to as the SMP set moving forward. Our experiment aims to recommend opinions from the PAR and SMP sets that could lead to a higher MPP.

4 Our Approach

Our proposed method utilizes a three-step approach which incorporates two concepts to evaluate the given opinion: strength degree estimation and argument-based analysis. Initially, we estimate the strength degree of each sentence within the given opinion. Subsequently, we categorize each sentence as claim, premise, or others. Following this, we identify the relationships between claims and premises, thereby discerning which premise supports a given claim and determining the strength

[6] https://www.ptt.cc/bbs/Stock/index.html

degree of the arguments. Finally, we recommend opinions based on the extracted strength degree and argument features. Detailed explanations of our proposed method are provided in this section.

4.1 Strength Degree

To acquire the strength degree (SD), we suggest using the difference between the price target (PT) and the most recent close price as a proxy. Equation 3 formulates our method for obtaining SD.

$$SD = \frac{PT - C_{t^r}}{C_{t^r}} \tag{3}$$

Here, C_{t^r} is the close price at the time the report was released on date t^r.

For generating a dataset with FSD, we collate an additional 628 professional analysis reports from Bloomberg Terminal. These reports do not overlap with the ones introduced in Section 3.2. We extract the PT, C_{t^r}, and the content from these reports. As a result, we obtain 37,147 sentences from these reports, with sentences from the same report being labeled with the same SD.

These sentences are segregated into two groups based on whether the SD of the sentence is higher than the average SD (22.51%) of all sentences or not. We then fine-tune BERT-Chinese (Devlin et al., 2019) to perform this binary classification task. Using 80% of the instances for training and the remaining for testing, we develop a model that achieves micro- and macro-averaged F-scores of 66.78% and 62.10%, respectively. This fine-tuned model is then employed to predict each sentence in the PAR and SMP sets. We use the probability that the given sentence will have SD higher than the average as the fuzzy strength degree (FSD), which ranges from 0 to 1. Hence, all sentences in PAR and SMP sets are assigned an FSD.

Table 1 displays some instances and the corresponding FSD. We observe that a higher FSD value tends to more strongly support a bullish stance. This observation suggests that the proposed approach can effectively estimate the strength degree.

FSD	Sentence
0.93	Expected to grow in both old and new businesses
0.88	Driven by memory and Indian factories
0.85	Profits in 2019 will show explosive growth
0.59	the company implements epidemic prevention measures
0.50	companies are expected to introduce new products
0.15	because of Chinese manufacturers bidding for orders
0.04	Considering the slow recovery speed of operations
0.01	Operation still hasn't got rid of the downturn

Table 1: Sentences and the corresponding FSD.

4.2 Argument-based Opinion Analysis

Argument mining has been a major focus of research over the past decade (Lawrence and Reed, 2020; Cabrio and Villata, 2018). Fundamental tasks such as argument detection are crucial in analyzing narratives like debate discourse and persuasive essays (Shnarch et al., 2020; Levy et al., 2018; Shnarch et al., 2018; Rinott et al., 2015; Chen et al., 2020b). Furthermore, the linking of arguments, claims, and premises can form a structure that provides a better understanding of the author's strategy and narrative flow (Li et al., 2020; Beigman Klebanov et al., 2016; Wachsmuth et al., 2016; Chen et al., 2021). Building on these previous works, we incorporate both argument detection and argument relation linking tasks in our analysis of investors' opinions.

We utilize the pretrained BERT-Chinese models proposed by Chen et al. (2021) for argument detection and argument relation linking. These models yield macro-averaged F-scores of 79.86%, 57.69%, and 56.96% for claim detection, premise detection, and relation linking, respectively, on the Equity-AMSA dataset (Lin et al., 2024). We use these models to categorize the sentences in the PAR and SMP datasets as either claims or premises. The pretrained relation linking model is then used to identify whether a sentence designated as a premise supports an identified claim.

This approach allows us to dissect an opinion across multiple dimensions. For instance, consider the sentences "Profits in 2019 will show explosive growth" and "because of Chinese manufacturers bidding for orders" from Table 1. While the former is a claim and the latter a premise, do they serve the same function in shaping investors' opinions? Given that the claim predicts future earnings or operations and the premise describes occurred or plausible events, we aim to determine which part should receive more focus in recommending investors' opinions. In addition, we aim to investigate whether sentences not related to the claim (like greetings) should be included or excluded from the analysis.

In our experiments, we explore these research questions using the assigned argument labels and propose the following strategies to evaluate the strength degree of each opinion in the PAR and SMP datasets:

- *AllSent*: Average the FSD of all sentences in the opinion.
- *AllArg*: Average the FSD of the sentences labeled as either a claim or a premise.
- *ClaimOnly*: Average the FSD of the sentences labeled as a claim.
- *PremiseOnly*: Average the FSD of the sentences labeled as a premise.
- *KeyPremise*: Use the maximum FSD among all premises linked to the claims.

In summary, we investigate various strategies to aggregate sentence-level FSD to opinion-level FSD and rank investors' opinions based on this opinion-level FSD.

5 Experiment

5.1 Baselines

We compare our approach against several strong baselines from prior work.

Following Zong et al. (2020), we first implement a series of BERT-based models that predict the profitability of analyst reports. *BERT-Conf* classifies reports into higher MPP (Top 50%) and lower MPP (Bottom 50%) groups using BERT, and uses the probability of belonging to the top group as the ranking score. *BERT-Reg* treats MPP prediction as a regression task, ranking reports by the predicted continuous MPP value. *Mengzi-FinBERT-Reg* replaces BERT in BERT-Reg with Mengzi-FinBERT (Zhang et al., 2021), a model pre-trained on financial news and analyst reports. These baselines serve as standard BERT-derived content-based ranking methods.

Chen et al. (2021) proposed *ExpertLike*, which counts the number of sentences in each social media post that are classified as professionally written and ranks posts accordingly. Because only 23.16% of posts in the SMP dataset contain at least one expert-like sentence, many posts receive zero scores. We also test a variant, *ExpertLike + FSD*, which averages our fuzzy strength degree (FSD) values over expert-like sentences to provide a more fine-grained ranking signal.

Chen et al. (2024b) introduced a professionalism-aware *pre-finetuning* (PF) framework between pre-training and task fine-tuning to enhance profitability ranking. The PF stage contains two auxiliary objectives: (i) sentence-level (SLPF) and (ii) word-level (WLPF). Following their reported best configurations, we use *SCQF + WLPF* for analyst reports (PAR) and *SCQF + SLPF* for social media posts (SMP). Here SCQF denotes "Sbert-Chinese-QMC-Finance-v1" pretrained language model. When rank-based evaluation is required, we adopt their pairwise ranking setup, which they showed outperforms regression for profitability prediction. This professionalism-aware PF serves as a strong initialization baseline for our argument-based method.

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
BERT-Conf (Zong et al., 2020)	11.68%	12.42%	15.02%	15.14%
BERT-Reg (Devlin et al., 2019)	12.96%	12.58%	13.23%	12.05%
Mengzi-FinBERT-Reg (Zhang et al., 2021)	10.97%	12.08%	14.96%	15.29%
SCQF + WLPF (Chen et al., 2024b)	14.62%	**15.97%**	20.57%	10.67%
AllSent	15.25%	14.53%	12.75%	11.93%
AllArg	14.36%	14.75%	12.73%	11.93%
ClaimOnly	14.27%	14.51%	12.78%	11.39%
PremiseOnly	14.51%	14.35%	9.39%	2.46%
KeyPremise	**15.59%**	14.71%	*5.01%*	*1.46%*

Table 2: Averaged MPP of the sorted analyst reports. For the top-ranked reports, we bold the highest one; and for the ranked-last reports, we use italic to mark the worst one.

5.2 Recommendation Results in PAR set

For a fair comparison, we incorporate the additional 628 reports discussed in Section 4.1 in the training process. Additionally, the method proposed by Chen et al. (2021), which is based on the principle of counting expert-like sentences in social media posts, is unsuitable for ranking analyst reports because all sentences in the PAR set are expert-written.

We adopt the evaluation method in Chen et al. (2021) to report the average MPP in the 10th and 9th deciles. We also report the last two deciles to demonstrate whether the proposed method can highlight opinions with higher MPP and filter out those with lower MPP. Table 2 displays the average MPP of the top-ranked and lowest-ranked reports based on different strategies. A Friedman test indicates significant differences across strategies (p = 0.009).

Several findings can be drawn from Table 2: First, the baseline model, BERT-Conf (Zong et al., 2020), appears ineffective for ranking reports as the reports with a higher probability in the top 50% group yield a lower MPP. For a more detailed analysis, we calculate the average MPP of all opinions categorized as the top 50% and bottom 50% groups, comparing the results with the best-performing strategy, *KeyPremise*. Second, we observe a correlation between the proposed FSD and potential profit (MPP). Regardless of the strategies adopted, the top-ranked opinions consistently lead to higher potential profit than the lowest-ranked opinions. This implies that the proposed FSD can be effectively used to rank investors' opinions. Third, the ordering results of *AllSent*, *AllArg*, and *ClaimOnly* are strikingly similar in both top-ranked and lowest-ranked opinions. Both *PremiseOnly* and *KeyPremise* effectively filter out the lowest-ranked opinions. The only difference between *PremiseOnly* and *AllSent* is that *PremiseOnly* exclusively uses the FSD of premises to order the opinions. This result underscores the importance of the premise in ranking investors' opinions. We also observe an improvement in filtering out opinions with lower MPP when only the key premise is used.

We also compare against the professionalism-aware pre-finetuning baseline of Chen et al. (2024b). In Table 2, *SCQF + WLPF* achieves the best 9th-decile MPP (15.97%), slightly outperforming our *KeyPremise* (14.71%) and *AllSent* (14.53%). This indicates that injecting professionalism signals before task training can surface strong mid-top reports even without explicit argument structure. By contrast, our *KeyPremise* yields the *best* 10th-decile MPP (15.59%) and, more importantly, is substantially better at filtering low-quality items: its last two deciles drop to 5.01% and 1.46%, whereas *SCQF + WLPF* remains relatively high (20.57% and 10.67%). In short, our argument-based ranking is more effective for trimming the bottom tail.

5.3 Recommendation Results in SMP Set

Unlike previous research (Zong et al., 2020; Chen et al., 2021), which focused either on analyst reports or social media posts, our study explores both realms. In this section, we employ ExpertLike (Chen et al., 2021) as the benchmark. ExpertLike ranks social media posts by counting the number of sentences in the posts that are classified as professionally written. Their method, which sorts social media posts based on professional-appearing content, has been demonstrated to be profitable.

Building on their approach, we propose an alternative strategy for comparison: *ExpertLike + FSD*: Compute the average of the FSD of the sentences classified as expert-like.

Due to the absence of an additional dataset with MPP labels, we cannot employ supervised models in this section. That is, the lack of argument annotations for social media posts precludes testing *AllArg*, *ClaimOnly*, *PremiseOnly*, and *KeyPremise* strategies. Although we attempted to use the pretrained models from Chen et al. (2021) for professional report argument analysis, none of the sentences in the SMP set were identified as claims or premises. Thus, we apply the *AllSent* strategy in our experiment, examining whether the proposed FSD also applies to amateur investors' opinions.

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
ExpertLike + FSD	16.18%	12.98%	-	-
ExpertLike (Chen et al., 2021)	17.61%	13.09%	-	-
SCQF + SLPF (Chen et al., 2024b)	**23.39%**	11.60%	*11.91%*	*7.47%*
AllSent	19.41%	**15.79%**	12.38%	9.93%

Table 3: Averaged MPP of the sorted amateur posts based on argument strength.

Table 3 presents the experimental results in the SMP set. Firstly, we find that the top-ranked posts sorted by argument strength yield higher profits than those sorted by ExpertLike-based counting strategies. Secondly, the *ExpertLike + FSD* results indicate that the fusion of previous work and the proposed method successfully ranks the opinions, although the potential profits using this combined strategy are not as high as implementing the two approaches separately. Thirdly, we observe that only 23.16% of posts in the SMP set contain at least one expert-like sentence as per the ExpertLike approach (Chen et al., 2021). Consequently, the majority of posts (76.84%) receive zero scores when using the method proposed in the previous work, which restricts the applicability of their approach in ranking opinions. In contrast, the *AllSent* results demonstrate that our proposed method can also be applied to rank social media posts, reinforcing the correlation between the proposed FSD and the MPP of amateur opinions.

However, the professionalism-aware pre-finetuning baseline *SCQF + SLPF* (Chen et al., 2024b) achieves the highest top-decile MPP (23.39%) and lowest last-decile scores (11.91%, 7.47%) . These results suggest that, for short and noisy social media texts, surface-level cues such as wording and professionalism already capture most of the informative signal, and explicit argument structures contribute less. In contrast, our approach shows clearer advantages on professional analyst reports (PAR), where well-formed claims and premises, along with quantified opinion strength (*KeyPremise*/*AllSent*), align more directly with profitability-oriented ranking. Another plausible explanation is that, as noted by Chen et al. (2020a), financial social media posts—especially very

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
BERT-Conf (Zong et al., 2020)	-10.59%	-10.44%	-10.38%	-10.05%
BERT-Reg (Devlin et al., 2019)	-10.40%	-10.39%	-10.58%	-11.55%
Mengzi-FinBERT-Reg (Zhang et al., 2021)	-9.70%	**-9.95%**	-11.52%	-11.42%
SCQF + WLPF (Chen et al., 2024b)	-13.91%	-12.62%	-11.93%	-13.16%
AllSent	-11.30%	-11.98%	-9.24%	-8.80%
AllArg	-11.48%	-10.81%	-10.30%	-10.39%
ClaimOnly	-10.70%	-10.68%	-10.75%	-10.50%
PremiseOnly	-10.53%	-10.95%	-12.22%	-19.95%
KeyPremise	**-9.47%**	**-9.95%**	*-15.45%*	*-22.53%*

Table 4: Averaged ML of the sorted analyst reports. Because lower is better for risk (ML), we bold the lowest values for the top-ranked results and italicize the highest values for the last-ranked results.

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
ExpertLike (Chen et al., 2021)	-3.72%	**-6.26%**	-	-
ExpertLike + FSD	-4.23%	-6.38%	-	-
SCQF + SLPF (Chen et al., 2024b)	**-1.68%**	-7.76%	-7.55%	-8.74%
AllSent	-10.22%	-8.33%	-5.41%	-7.42%

Table 5: Averaged ML of the sorted amateur posts.

short ones—rarely include explicit reasoning or evidence, making argument-based methods less applicable to SMP.

6 Discussion

6.1 Risk Analysis

Risk consideration is a vital aspect of financial decision-making. Proper understanding of associated risks can guide investors in selecting strategies that align with their risk tolerance. In this section, we examine the risk associated with the top-ranked and the last-ranked opinions. We employ the maximum loss (ML) metric, as proposed by prior work (Chen et al., 2021), to evaluate the risk. ML captures the potential maximum loss if one were to trade following a particular opinion. The ML of bullish and bearish reports are calculated using the following equations:

$$ML_{bullish} = \min_{t+1 \leq i \leq t+T} \frac{L_i - O_{t+1}}{O_{t+1}} \tag{4}$$

$$ML_{bearish} = \max_{t+1 \leq i \leq t+T} \frac{O_{t+1} - H_i}{O_{t+1}} \tag{5}$$

Table 4 presents the MLs under various strategies. Interestingly, the opinions sorted by the *KeyPremise* not only yield higher profits but also pose lower risks. This pattern is observed in both the top-ranked and last-ranked scenarios, emphasizing the utility of the proposed method in investor opinion recommendation tasks.

From Table 5, the professionalism-aware baseline *SCQF + SLPF* (Chen et al., 2024b) yields the smallest losses (closest to zero) for the top-ranked SMP items. This indicates that surface professionalism/wording cues not only improve top-decile MPP but also reduce downside risk for high-ranked social posts. Overall, on short and noisy social media texts, *SCQF + SLPF* offers better risk control where it matters most (top ranks). Taken together with the MPP and ML results on SMP, our findings suggest that the proposed fine-grained, argument-centric analysis (claims/premises plus FSD aggregation) is *less effective* for short and noisy social media posts. In this setting, surface signals related to wording and professionalism—as captured by *SCQF + SLPF* (Chen et al., 2024b)—tend to dominate both upside (top-decile MPP) and downside control (top-decile ML), whereas explicit argument structure contributes limited additional benefits. By contrast, on professional analyst reports (PAR), where arguments are more complete and quantifiable, our method shows clearer advantages.

6.2 Behavior of Professional Analysts

The actions of professional analysts carry significant influence in the financial market, impacting future price movements and other investors' decisions. Conrad et al. (2006) examined professional analysts' reactions to public information such as news when making recommendations. They found that analysts' recommendations can help predict future price movements.Hirst et al. (1995) found that the strength of arguments in analysts' reports influence investors' judgments. Niehaus and Zhang (2010) explored how research coverage by analysts affects the market share of the broker. This topic has also been explored by Chen et al. (2024a), who examined professional behavior in investment contexts through hierarchical multi-agent simulations, analyzing how large-scale language models replicate analysts' and traders' decision patterns. Inspired by these studies, we aim to examine whether professional analysts and traders have different reactions towards top-ranked and last-ranked opinions.

Unlike previous studies (Zong et al., 2020; Chen et al., 2021) that focused solely on profitability and risk, we propose two novel discussion directions: (1) Do recommended analyst reports influence other professional analysts' views in the near future? (2) Do professional traders take actions (buy/sell/no action) in alignment with the recommendations (overweight/underweight/neutral) in the recommended analysts' reports? The analyses in this section aim to elucidate potential future outcomes after following the recommended opinions in trading.

To determine whether recommended analyst reports influence other professional analysts' views in the near future, we gather recommendation data from all analysts via the Bloomberg Terminal. We align the reports in the PAR set with the collected data based on the date. For each report in the PAR set, we determine whether any professional analyst changed their view on the same stock in the period from $t+1$ to $t+6$, where t is the report release date. We use P_{ANA}, as defined in Equation 6, to compare the reactions of professional analysts to recommended opinions across different strategies.

$$P_{ANA} = \frac{N^{f}_{RR_i}}{N_{RR_i}} \tag{6}$$

Here, RR_i represents the recommended reports in the ith decile, N_{RR_i} is the total number of recommended reports in the ith decile, and $N^{f}_{RR_i}$ represents the number of reports in RR_i that meet the criteria (i.e., at least one professional analyst changes their view on the same stock in the period from $t+1$ to $t+6$).

Table 6 provides the statistics of P_{ANA} for reports in different deciles, ranked by the *KeyPremise* strategy. Intriguingly, top-ranked reports are more likely to influence a change in other professional analysts' views compared to last-ranked reports. This finding implies that reports recommended by our proposed approach receive more attention from professional analysts than those at the bottom of the ranking.

	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
P_{ANA}	10.53%	10.75%	0.00%	0.00%

Table 6: Analysis of the professional analysts behaviors after the report released date. The recommendation in this table is based on *KeyPremise* strategy.

6.3 Professional Traders' Behaviors

Although everyone can express opinions, not all opinions will manifest in market actions. When investors put their money into a specific stock, it shows a strong affirmation of the said stock. Therefore, investors' trading behaviors have long been a focal point of research. Chordia et al. (2001a) examined the relationship between market liquidity and trading activities, discovering increased trading activity preceding macroeconomic information announcements. Furthermore, Chordia and Subrahmanyam (Chordia et al., 2001b) correlated trading activities with expected returns, asserting the importance of trading activity in the cross-section of expected returns. In another study, Wüstenfeld and Geldner (2021) explored Bitcoin trading activity, suggesting investors in different countries exhibit varied attitudes toward Bitcoin investment.

Unlike these previous studies, which focus on using market volume-related information to understand investor trading activities, our research targets the trading behaviors of professional traders—those employed in financial institutions. We hypothesize that if professional traders buy (or sell) a stock mentioned in a recommended bullish (or bearish) report, they are likely in agreement with the report's opinions.

To investigate whether professional traders concur with recommended opinions, we collected publicly available transaction records from the Taiwan Stock Exchange. These records contain details on the number of units that professional institutions buy/sell for a specific stock on a given date. The records are further categorized into three groups: Qualified Foreign Institutional Investors (QFII), Investment Trust (Fund), and Dealer. Using these records, we determined whether these professional traders traded stocks in the same direction as the recommended reports. We calculate the concurring ratio (CR) as defined in Equation 7.

$$CR = \frac{N^{c}_{RR_i}}{N_{RR_i}} \tag{7}$$

Here, $N^{c}_{RR_i}$ is the number of times professional traders act in the same direction as the reports in RR_i on the next trading day after the report release.

Table 7 presents the results in the PAR set. We observe that top-ranked opinions garner more favor from professional traders than ranked-last opinions, irrespective of institution type. As QFII's positions represent about 40% of the total stock market capitalization, their behaviors are considered more influential. Notably, QFII shows a high agreement towards recommended stocks. This suggests our proposed method successfully identifies important premises that align with professional traders' decision-making process.

To sum up, we utilize professional behaviors as an indicator of our proposed method's hit ratio, discussed in Section 6.2 and this section. Results

	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
CR-QFII	50.44%	50.66%	27.81%	0.88%
CR-Fund	34.65%	34.43%	7.51%	15.11%
CR-Dealer	39.04%	41.23%	9.27%	18.67%

Table 7: Analysis of the professional traders' behaviors after the report released date. The recommendation in this table is based on *KeyPremise* strategy.

indicate that our approach effectively identifies premises aligned with professional perspectives and filters out those that do not pique professional traders' interest.

6.4 Evaluation on Full Recommendation List

To perform a more granular evaluation in the investor opinion recommendation scenario, Table 8 presents the results of average MPP and ML of top-10 and top-20 reports, sorted by each method. *AllSent* performs the best under this setting, while *KeyPremise* ranks second in terms of MPP. From the risk perspective (ML), *AllSent* still outperforms other methods. These findings confirm that *AllSent*, which employs the proposed FSD to rank investor opinions, is the most suitable method for investor opinion recommendation tasks among all methods tested.

Firstly, despite *AllSent* slightly underperforming *KeyPremise* in MPP in the 10th decile (Top-228 Reports) and the 9th decile (Top-456 reports) as depicted in Table 2, the top-10 and top-20 reports recommended by *AllSent* achieve higher MPP and lower ML than those based on other methods. The real-world application scenario likely involves recommending 10 to 20 reports, allowing investors to review these suggestions and make investment decisions based on their experience. Secondly, although *AllSent* lags slightly behind *KeyPremise* in filtering out opinions leading to lower MPP as displayed in Table 2, the overall ranking results based on nDCG show that *AllSent* is successful in ranking opinions by MPP. Since investor opinion recommendation aims to identify the opinions leading to profitable outcomes, we propose that the *AllSent* strategy is ideally suited for this purpose. Thirdly, *AllSent* has proven to be applicable in both professional and amateur investors' opinion recommendations. However, *KeyPremise* requires additional annotation data and is currently unexplored in the context of amateur investor opinion recommendation.

	Top-10 MPP	Top-20 MPP	Top-10 ML	Top-20 ML
BERT-Conf (Zong et al., 2020)	8.84%	7.38%	-11.53%	-14.11%
BERT-Reg (Devlin et al., 2019)	15.91%	12.94%	-12.94%	-12.60%
Mengzi-FinBERT-Reg (Zhang et al., 2021)	13.16%	9.90%	-8.28%	-9.98%
SCQF + WLPF (Chen et al., 2024b)	16.50%	16.29%	-11.93%	-12.79%
AllSent	**18.89%**	**17.69%**	**-4.05%**	**-6.65%**
AllArg	9.04%	12.70%	-10.73%	-10.87%
ClaimOnly	13.28%	13.47%	-11.16%	-9.61%
PremiseOnly	11.01%	12.76%	-6.31%	-6.62%
KeyPremise	14.75%	15.44%	-12.71%	-11.72%

Table 8: MPP and ML of Top-10 and Top-20 posts.

7 Conclusion

We presented an argument-centric framework for investment opinion recommendation that combines a fuzzy strength degree (FSD) with claim/premise detection and linkage. On professional analyst reports (PAR), our methods are consistently effective: (i) they improve top-decile profitability (e.g., *KeyPremise*) and (ii) more aggressively filter the lower tail (low-MPP deciles) while maintaining favorable risk profiles (lower ML). The behavioral analyses further show that top-ranked reports under our approach are more likely to be followed by view changes from other analysts and to align with subsequent trades by professional institutions, indicating practical relevance beyond offline metrics. On social media posts (SMP), however, the picture is different. This suggests that surface signals related to wording and professionalism can dominate in SMP, whereas explicit argumentative structure is most beneficial when arguments are well-formed and quantifiable (as in PAR).

Overall, our findings indicate a domain-sensitive strategy: use argument-based strength modeling to rank and de-risk analyst reports, and rely more on professionalism-aware signals for social media posts. An immediate avenue for future work is a hybrid system that adaptively combines professionalism-aware pre-finetuning with argument-strength features, selecting the mixture by domain or even per-document cues. Beyond that, we plan to extend to multilingual and cross-market settings, and study human-in-the-loop ranking that balances upside potential with risk constraints in different investor profiles.

Additionally, in light of recent discussions on report generation by large language models (Goldsack et al., 2025; Takayanagi et al., 2025), our proposed framework and evaluation protocol may also serve as a foundation for assessing the quality of generated investment reports. Future work can explore how argument structure and strength modeling contribute to the automatic evaluation of agent-generated analyses.

Limitation

First, our datasets are predominantly one market. This geographical and linguistic limitation might affect the generalizability of our findings to other regions and languages. Further studies should consider incorporating datasets from diverse markets and languages to validate the robustness and applicability of our approach globally. Second, our analysis of professional analysts' and traders' behaviors is based on publicly available transaction records and recommendation data. However, these records might not capture the full spectrum of professional activities and decision-making processes. More granular data, including intraday trading patterns and proprietary analyst notes, could offer deeper insights into professional behaviors in response to recommended opinions.

Ethical Considerations

The use of systems in financial markets raises ethical concerns, particularly regarding market manipulation and the amplification of biased opinions. It is crucial to implement safeguards and ethical guidelines to ensure that the system promotes fair and unbiased recommendations, preventing any potential misuse.

Acknowledgments

This work was supported in part by AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain." It was also supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2634-F-002-003- and 114-2221-E-002-070- MY3, and Ministry of Education (MOE), Taiwan, under grant NTU-114L900901.

References

Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. 2011. Identifying and following expert investors in stock microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75, Berlin, Germany. Association for Computational Linguistics.

Johan Bollen and Huina Mao. 2011. Twitter mood as a stock market predictor. *Computer*, 44(10):91–94.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020a. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110, Marseille, France. European Language Resources Association.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020b. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, page 1973–1976, New York, NY, USA. Association for Computing Machinery.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, WWW '21, page 3987–3998, New York, NY, USA. Association for Computing Machinery.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024a. Hierarchical organization simulacra in the investment sector. *arXiv preprint arXiv:2410.00354*.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024b. Professionalism-aware pre-finetuning for profitability ranking. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3674–3678.

Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. 2001a. Market liquidity and trading activity. *The Journal of Finance*.

Tarun Chordia, Avanidhar Subrahmanyam, and V Ravi Anshuman. 2001b. Trading activity and expected stock returns. *Journal of Financial Economics*.

Jennifer Conrad, Bradford Cornell, Wayne R Landsman, and Brian R Rountree. 2006. How do analyst recommendations respond to major news? *Journal of Financial and Quantitative Analysis*, 41(1):25–49.

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.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2025. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.

D Eric Hirst, Lisa Koonce, and Paul J Simko. 1995. Investor reactions to financial analysts' research reports. *Journal of Accounting Research*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Chin-Yi Lin, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Argument-based sentiment analysis on forward-looking statements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13804–13815, Bangkok, Thailand. Association for Computational Linguistics.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22.

Greg Niehaus and Donghang Zhang. 2010. The impact of sell-side analyst research coverage on an affiliated broker's market share of trading volume. *Journal of Banking & Finance*.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697, Online. Association for Computational Linguistics.

Takehiro Takayanagi, Tomas Goldsack, Kiyoshi Izumi, Chenghua Lin, Hiroya Takamura, and Chung-Chi Chen. 2025. Earnings2Insights: Analyst report generation for investment guidance. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 246–251, Suzhou, China. Association for Computational Linguistics.

Wenting Tu, Min Yang, David W Cheung, and Nikos Mamoulis. 2018. Investment recommendation by discovering high-quality opinions in investor based social networks. *Information Systems*, 78:189–198.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Jan Wüstenfeld and Teo Geldner. 2021. Economic uncertainty and national bitcoin trading activity. *The North American Journal of Economics and Finance*.

Wei Xiao. 2015. Does practice make perfect? evidence from individual investors' experiences and investment returns. *Journal of Interdisciplinary Mathematics*, 18(6):811–825.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

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.