Evaluating the Rationales of Amateur Investors

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Department of Computer Science, National Chengchi University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@nccu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

Social media's rise in popularity has demonstrated the usefulness of the wisdom of the crowd. Most previous works take into account the law of large numbers and simply average the results extracted from tasks such as opinion mining and sentiment analysis. Few attempt to identify high-quality opinions from the mined results. In this paper, we propose an approach for capturing expert-like rationales from social media platforms without the requirement of the annotated data. By leveraging stylistic and semantic features, our approach achieves an F1-score of 90.81%. The comparison between the rationales of experts and those of the crowd is done from stylistic and semantic perspectives, revealing that stylistic and semantic information provides complementary cues for professional rationales. We further show the advantage of using these superlative analysis results in the financial market, and find that top-ranked opinions identified by our approach increase potential returns by up to 90.31% and reduce downside risk by up to 71.69%, compared with opinions ranked by feedback from social media users. Moreover, the performance of our method on downside risk control is comparable with that of professional analysts.

CCS CONCEPTS

• **Information systems → Document filtering**.

KEYWORDS

rationale evaluation, social trading, opinion quality

ACM Reference Format:
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), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449964

1 INTRODUCTION

When expressing opinions, people not only indicate what to do or what not to do, but also provide the rationale behind their viewpoints. The persuasiveness of this rationale influences the outcome of their opinions. In this paper, we attempt to mine high-quality opinions by inspecting the supporting rationales.

Free-form posts on social media platforms have increased rapidly from the early 21st century. Many researchers have focused on extracting opinions from social media and using them for various applications. However, crowd opinions are not always useful because of the inherent noise in the posts. Thus, a fine-grained analysis is needed to evaluate the quality of crowd opinions. The approaches can be roughly categorized into two types.

(1) **Opinion classification based on post content**: This type of approach classifies a post as *useful* or *useless*. It is expensive to annotate a sufficient number of training instances for different domains.
(2) **Opinion classification based on reader feedback**: This type of approach leverages reader information such as the number of Facebook likes. However, this kind of approach cannot predict the quality of a post that was just published, as no reader information is available yet. It is also difficult to evaluate posts from a new account or from accounts with few followers. Furthermore, Twitter, Facebook, and Instagram plan to hide information about likes. The above issues all decrease the feasibility of this kind of approaches in future applications.

In this paper, we address issues of annotation cost, cold starts, and few followers. We conduct experiments that show the usefulness of the proposed approach. The methodology can be extended into various application scenarios.

Our approach is based on a simple idea: high-quality posts from the crowd may share characteristics with articles written by experts. Therefore, in this paper, we use documents written by experts and the crowd as our dataset, and train models to discriminate expert rationales from the crowd's rationales. Leveraging the high accuracy of the models, we further use the outcomes of the model to mine high-quality opinions from the crowd. That is, if rationales written by the crowd are predicted to be expert rationales, we infer that the quality of the opinions in these documents is higher than that of the opinions in documents predicted to be the crowd's rationales. In this paper, we evidence the usefulness of this approach in the financial application scenario, and experiment on the financial analysis reports from both professional analysts (experts) and amateur investors (crowd).

Inspired by Basile et al. [2], who show that stylistic information of posts is useful for predicting information about the writer such as social stratification, we explore both stylistic and semantic features

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '21, April 19–23, 2021, Ljubljana, Slovenia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449964

to identify expert-like rationales. In Section 5, we discuss in detail the different cues provided by both features.

In this paper, we attempt to answer the following research questions:

- **(RQ1)** To what extent can we use stylistic and semantic features to differentiate between rationales from professional analysts and amateur investors?
- **(RQ2)** If we are able to classify rationales successfully, which kind of features is more useful?
- **(RQ3)** Which approach is better, following high-quality opinions mined by the proposed approach, or following opinions ranked according to the feedback of social media users?

We will answer **(RQ1)** and **(RQ3)** in Section 4, and discuss **(RQ2)** in Section 5. Experimental results provide positive and impressive findings toward **(RQ3)**.

For the experimental setup, we collect more than 70K sentences related to trading rationales from reports of professional analysts and from posts of amateur investors on social media platforms. Following previous work [2], we use a dependency tree and part-of-speech (POS) tags to represent stylistic information, and adopt BERT pre-trained embeddings [12] to represent semantic information.

Our contributions are listed as follows.

(1) We propose a novel method to infer the persuasiveness of rationales, and further use these results to mine high-quality opinions from the crowd.
(2) We provide the pioneer results in identifying expert-like rationales from financial social media data.
(3) We explore a new direction in using crowd opinions, and show clear differences between high-quality and low-quality opinions from both profit and risk aspects.

The rest of this paper is organized as follows. In Section 2, we give a survey on the previous works and compare them with our proposed approach. In Section 3, we illustrate our task settings and present statistics of the dataset for differentiating rationales from professional analysts and from amateur investors. In Section 4, we explore the approaches for rationale classification and further compare the outcomes of opinions of different qualities. In Section 5, we provide an in-depth discussion of the results of our approach, and compare both stylistic and semantic information of rationales from experts and the crowd. In Section 6, we further propose a new dataset for future work to probe a novel extended task, claim-rationale inference. In Section 7, we list some research directions for future works based on the notions of argument mining. We conclude this paper in Section 8.

2 RELATED WORK

In the last two decades, many works have focused on user-generated content on the Internet, including blogs [35, 38], online forums [15, 39], e-commerce platforms [1], and social media [14, 28]. Most of them attempt to mine opinions from various types of textual data [16, 20, 27]. In contrast, only a few works attempt to evaluate opinion quality. One related topic already explored is the helpfulness of online product reviews. Ocampo Diaz and Ng [25] provide a survey on advances in this regard. The ground truth of the dataset is labeled by users on e-commerce platforms such as Amazon.com. As this series of research focuses solely on e-commerce platforms, it cannot be extended to other sites such as social media and other user-generated text easily. As suggested in Ocampo Diaz and Ng [25], future work should be explored on other platforms. In this paper, we provide a general approach to mine high-quality opinions, and show that the proposed approach can successfully mine the high-quality opinions from financial social media platforms. In the future, the proposed approach can be extended to various application scenarios such as fake news detection and writing evaluations in education.

Feature-based methods have been developed to rank argumentative comments [36] and product reviews [13]. To the best of our knowledge, few works explore the quality of the rationales, i.e., the reasons supporting their opinions, to rank the opinions. This work is the first attempt to mine high-quality opinions via the persuasiveness of the rationales.

Ying and Duboue [40] annotate a pilot dataset and classify rationales into four levels for educational purposes. In contrast to their work, which simply uses a vanilla neural network model with semantic information directly, we investigate both stylistic and semantic information and show the importance of stylistic cues to capture expert-like rationales. Furthermore, our novel approach does not require labeled data. In this paper, we provide an in-depth discussion of the applications of the captured expert-like rationales.

Financial social media data is a recent focus of researchers in natural language processing and in the financial domain. Chen et al. [6] propose numeral attachment, a new task to capture the relation between cashtags and numerals in financial tweets. Lin et al. [24] use sentiment on social media platforms to predict company sales. Xu and Cohen [37] adopt both tweets and market prices to predict stock movement. Few previous works on financial social media data attempt to evaluate or rank investor opinions. The experimental results in this paper show the returns following the top 10% of opinions on financial social media platforms are 47.49% greater than the returns following non-expert-like opinions. That supports the necessity of opinion quality assessment.

Recently, some works explore the uses of the style of free-form posts to extract information about the post writers. Basile et al. [2] find that the stylistics of restaurant reviews can indicate the writer's social stratification. Zhang et al. [41] show that both writing and photography styles can be used to identify the drug trafficker. In this paper, we find that writing style yields opinions with the lowest risk for trading. Our experiments show that high-quality opinions characterized by an expert-like writing style reduce up to 85.76% of downside risk.

3 TASK SETTING AND DATASET

3.1 Classification and Quality Evaluation

In this paper, we first postulate that the rationales of experts are credible rationales, and further attempt to capture expert-like rationales from the crowd. In other words, if a rationale from the crowd is classified as an expert's rationale, either the style or the wording of the rationale is similar to that of an expert. We further infer that opinions supported by such expert-like rationales are of high quality. In Section 4, we present evidence supporting our postulation.

Given the above rationales, we first train the models to classify rationales from professional analysts and amateur investors on financial social media platforms. Second, we use the outcome of the

Table 1: Dataset statistics.

	Analyst	Crowd
Unique characters	2,737	3,298
Unique tokens	15,696	27,474
Unique POS tags	49	53
Unique tag-tag-arcs	415	622
Unique incoming arcs	25	25
Training set (sentences)	32,000	32,000
Test set (sentences)	812	812

best-performing model to evaluate the quality of amateur investors' opinions. That is, the more expert-like sentences in their posts, the higher quality their posts (opinions) are.

3.2 Dataset

We collect analysts' reports written in Chinese from Bloomberg Terminal[1], and parse one of the largest financial social media platforms in Taiwan, PTT Stock[2]. In analyst reports, analysts always use a subsection title to indicate key points in the subsection. In the content of the subsection, they present the rationales that support these key points. Thus, we extract the subsection content as the rationales of the analyst. In the social media platform, users follow a template to present their rationales and opinions about certain targets. Posts that do not follow this template are deleted by the administrator of the platform. Therefore, we use content between the "3. Analysis" and "4. Enter/Exit Strategies" subtitles as the rationales of amateur investors.

We further separate the content into sentences. Table 1 presents the statistics of the dataset. We adopt the Stanford dependency parser [9] to parse sentences, and use the POS tags of words and the head with the incoming arc (tag-tag-arc) to represent the dependency information. We also evaluate the performance using only incoming arcs (arc). Note that we balance the instances of both analysts and amateur investors to avoid data imbalance. As shown in Table 1, the wordings of analysts and amateur investors vary widely, but the stylistic features are similar. Note that all incoming arc features are the same for both analysts and amateur investors.

4 APPROACH

4.1 Discriminating Expert Rationales

To represent the features, we use the skip-gram method to pre-train word-level features, including 15-dimensional arc embeddings, 50-dimensional tag-tag-arc embeddings, 30-dimensional POS embeddings, and 300-dimensional word embeddings. To represent the character-level features, we use BERT (*bert-base-chinese*) [12], a pre-trained Chinese sentence encoder which provides 768-dimensional character embeddings.

We adopt convolutional neural network (CNN) [18] and bidirectional gated recurrent units (BiGRU) [10] for word-level features (dependency, POS, word token) to evaluate the performance on the task of discriminating rationales. For character-level features, we use the output embeddings of BERT as the input for both CNN and BiGRU models. They are named as BERT-CNN and BERT-BiGRU, respectively. We use the Adam optimizer [19] in our models. To avoid overfitting, we use a dropout layer with a 0.3 dropout rate and early stopping with a patience setting of 5 epochs.

[1]https://www.bloomberg.com/professional/solution/bloomberg-terminal/
[2]https://www.ptt.cc/bbs/Stock/index.html

Table 2: Experimental results discriminating analysts' rationales and amateur investors' rationales. (Dep. and TTA denote dependency and tag-tag-arc, respectively. * denotes results that are significantly different from the BERT-BiGRU model with character-level features under McNemar's test with $p < 0.05$.)

Features	Model	Macro-F1
Stylistic		
Dep. - arc	CNN	62.07
	BiGRU	61.54
Dep. - TTA	CNN	61.04
	BiGRU	62.91
POS	CNN	70.16
	BiGRU	73.34
Semantic		
Word-level	CNN	85.24
	BiGRU	85.74
Character-level	BERT-CNN	87.87
	BERT-BiGRU	88.59
Fusion Models		
BERT-BiGRU + BiGRU (POS) + BiGRU (TTA)		90.32
BERT-BiGRU + BiGRU (POS) + CNN (arc)		**90.81***

Fusion models are also constructed based on the experimental results of the single input models. We concatenate the output of each best-performing layer for different features and add a hidden layer and a dropout layer to optimize the model parameters. Finally, the softmax activation layer outputs the probability of each class. The macro-averaged F1 score is adopted as the evaluation metric.

Table 2 shows the experimental results of different models. The fusion model with incoming arc, POS, and character-level features performs the best, and achieves 90.81% in macro-F1 evaluation metric. Models with semantic features beat models with stylistic features by more than 10% of marco-F1. In most cases, the BiGRU architecture performs better than the CNN architecture.

The experimental results provide a positive answer to **(RQ1)**. Models discriminate expert rationales via both stylistic and semantic features with high F1-scores. More discussion on the stylistic and semantic features will be provided in Section 5 in an answer to **(RQ2)**.

Figure 1 illustrates the reason for the high performance of the semantic features. Wording varies widely between analysts and amateur investors. Only 20.35% of words are used by both analysts and amateur investors. Table 3 shows the prediction results of the best-performing model. Only 10% of the rationales of amateur investors are considered expert-like rationales. Based on the model predictions, we infer that opinions supported by these top-10% rationales are high-quality opinions. In the next subsection we present supporting evidence for this with an empirical study.

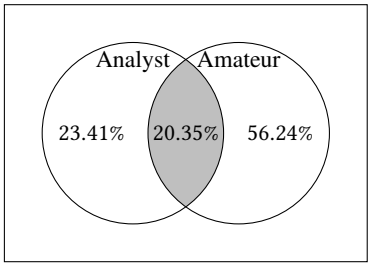

Figure 1: Venn diagram of wordings.

Table 3: Prediction results of best-performing model.

		Prediction	
		Analyst	Amateur investor
Actual	Analyst	0.92	0.08
	Amateur investor	0.10	0.90

4.2 Mining High-Quality Opinions

For the following empirical studies, we collect a new dataset consisting of posts on the same financial social media platform from 2019/05/13 to 2019/06/18. Note that there are no overlaps between this new dataset and the dataset used to train the discriminating models in Section 4.1. As described in Section 3.2, we use the sentences in the analysis section, i.e., content between "3. Analysis" and "4. Enter/Exit Strategies", as the expert-like rationale to score and rank post opinions. We use randomly-selected posts and posts ranked by the number of likes from other social media users as our baselines, and compare them with the ranking results of the best fusion model (Best FM), which use both stylistic and semantic features and the best model with stylistic features, i.e., the BiGRU model with POS features.

To calculate the maximum possible profit and the maximum loss as the performance of each post, we use the adjusted price, i.e., the prices from which we already remove the influence of corporate actions such as dividends, before 2019/09/11. It can better reflect the stock value and is commonly used in calculating the return. The adjusted price is computed by the data provider. During our backtesting period, the global market was influenced by the China-United States trade war. Thus, our results are more meaningful than those restricted to the long-term bullish market, especially in terms of the measurement of downside risk.

We enter the market at the opening price on day $t+1$ by following the opinion of the post on day t, and trace the maximum possible profit and the maximum loss during the backtesting period. In other words, we trace the unrealized return of the trading based on the opinions of amateur investors. We are able only to trace the unrealized return because most post authors present only opinions (bullish/bearish) and rationales, and do not specify when to exit the market, i.e., the timing to close their positions. For bullish opinions posted on day t, the maximum possible profit (MPP) and the maximum loss (ML) are calculated as

$$MPP_{bullish} = (\max(H_{(t+1,T)}) - O_{t+1})/O_{t+1} \quad (1)$$

Table 4: Performance of opinion quality ranking methods.

Method	Ranking	Average MPP	Average ML
Random		11.94%	-17.28%
Feedback	First decile	8.88%	-8.69%
	Second decile	7.14%	-10.73%
	Top 2 deciles	8.53%	-9.10%
Best FM	First decile	**17.61%**	-3.72%
	Second decile	8.80%	-8.67%
	Top 2 deciles	13.09%	-6.26%
BiGRU(POS)	First decile	15.78%	**-2.46%**
	Second decile	10.52%	-8.72%
	Top 2 deciles	12.71%	-6.11%

$$ML_{bullish} = (\min(L_{(t+1,T)}) - O_{t+1})/O_{t+1} \quad (2)$$

where O_t denotes the opening price of day t, $H_{(t,T)}$ denotes a list of the highest price of day t to day T, $L_{(t,T)}$ denotes a list of the lowest prices of day t to day T, and T is the last day of the backtesting period. For bearish opinions posted on day t, the MPP and the ML are calculated as

$$MPP_{bearish} = (O_{t+1} - \min(L_{(t+1,T)}))/O_{t+1} \quad (3)$$

$$ML_{bearish} = (O_{t+1} - \max(H_{(t+1,T)}))/O_{t+1}. \quad (4)$$

MPP sheds light on the potential profit, and also indicates the potential of the selected opinions. ML, in turn, provides information about the downside risk. We can use ML to determine whether the opinion was posted at the right time, i.e., whether bullish (bearish) opinions were posted at relatively lower (higher) price levels of the target financial instrument.

Table 4 shows the performance of different ranking methods.[3] Note that, we check the top 20% of amateurs' posts of different ranking methods manually to make sure the writer's opinion (bullish/bearish). Compared with randomly-selected crowd opinions without any expert-like rationales, the top-ranked opinions mined by our approaches are much better in terms of both evaluation metrics, in particular the averaged ML. On the other hand, the outcomes of our approaches also outperform the results of opinions ranked by the number of likes given by social media users.

To answer **(RQ3)**, we further separate opinions into more fine-grained groups. We find that regardless of the kind of ranking method adopted, the quality of the opinions ranked in the first decile is better than the quality of those in the second decile. This shows the benefit of using top-ranked crowd opinions.

In summary, a high-quality opinion not only provides a profitable suggestion, but also controls the downside risk. We propose an approach that yields better ranking results than approaches using only user feedback. Our approach can be used to evaluate the opinions of new posts as soon as they are published. Moreover, the proposed approach does not require annotated data.

[3] Some top-ranked posts do not contain opinions (bullish/bearish), i.e., they provide only an analysis of the market; thus the valid samples in the first decile and the second decile may not be equal, and the results of the top two deciles may not be equal to the average of the first and the second deciles.

Table 5: Readability comparison.

	Analyst	Crowd
Average hard words	**31.61**	24.60
Sentences with complex semantics	**6.86**	2.66
Noun phrase modifier ratio	**0.27**	0.16
Content word density	**0.87**	0.86
Positive transition words	**3.32**	1.98
Negative transition words	**0.99**	0.93
Number of personal pronouns	0.22	**0.98**
Number of negative words	0.11	**1.24**

Table 6: Selected words in expert-like lexicon.

Word	ELScore	Word	ELScore
Price target	2.10	Short	-1.72
Estimate	2.06	Guess	-1.75
We	2.04	I	-1.75
Gross profit ratio	1.91	Pattern	-1.71

5 DISCUSSION

5.1 Comparison with Analysts

Firstly, we use the Chinese Readability Index Explorer [33] to analyze the readability of the rationales of analysts and amateur investors. The statistics are shown in Table 5. We find that analysts use more difficult words than amateur investors, and also use more semantically complex sentences in their narratives. When reading through a noun phrase, readers must keep in mind the head of the noun phrase until they come to the modifier of the noun phrase. That is, using noun phrase modifiers can make sentences more complicated for readers. The noun phrase modifier ratio of analysts is 68.75% higher than that of amateur investors. Just and Carpenter [17] indicate that readers pay more attention to content words when reading. According to the statistics in Table 5, content word density is similar between the rationales of analysts and that of amateur investors.

Analysts tend to use more positive transition words when describing their rationales. The number of negative transition words is similar in the narratives of both analysts and amateur investors. We also find that amateur investors tend to use more personal pronouns when describing their rationales. Finally, analysts use few negative words in their reports, but amateur investors use many.

Secondly, in order to analyze the wordings of both analysts and amateur investors in-depth, we adopt the PMI measure as in previous work [23] to construct an expert-like lexicon, FinProLex.[4] The expert-like score (ELScore) of a word is calculated as

$$ELScore_w = \log_2 \frac{p(w, analyst)}{p(w)p(analyst)} - \log_2 \frac{p(w, amateur)}{p(w)p(amateur)}, \quad (5)$$

where w is the target word, *analyst* denotes analysts' reports, and *amateur* denotes the posts of amateur investors.

We list selected words to explain our findings in Table 6. Analysts provide rationales to support their "price targets", and amateur investors tend to provide rationales to support their views ("long" or "short"). That is, analysts evaluate stock values, and amateur investors attempt to predict price movement. "Gross profit ratio" and "pattern" indicate the different focuses of analysts and amateur investors. Analysts evaluate stock value based on fundamental analysis results, and amateur investors predict price movement using technical analysis such as technical indicators or chart patterns. The tone of analysts is more conscientious than that of amateur investors. For example, "estimate" yields a positive expert-like score and "guess" yields a negative expert-like score. We also find cues in personal pronouns such as "we" and "I".

[4] FinProLex: http://nlg.csie.ntu.edu.tw/nlpresource/FinProLex/

Table 7: Comparison of MPP and ML.

	Average MPP	Average ML
Analyst	**22.30%**	-6.52%
Stylistic + Semantic	17.61%	-3.72%
Stylistic	15.78%	**-2.46%**

Several works have noted the importance of numeral information in different domains [6, 30]. We calculate the numerals in the narratives of both analysts and amateur investors, and find that numerals occupy 12.53% and 7.68% of the space in analysts' reports and posts of amateur investors, respectively. This shows that analysts use more numerals as evidence to support their opinions.

Thirdly, we compare the MPP and ML results of analysts with top-ranked opinions mined by the proposed approach. According to Table 7, we find that although analysts identify targets with higher potential profit, the downside risk of trading based on analyst opinions is 2.65 times that of the downside risk of following top-ranked opinions of amateur investors. It shows that top-ranked opinions are comparable with the opinions of professional analysts.

In sum, in this subsection, we provide a discussion of the stylistic and semantic features of experts and the crowd in the financial domain. We also show the difference in the narratives between experts and the crowd. These insights can provide future research directions for work in other domains, when other researchers attempt to analyze expert opinions and opinions of the crowd. Furthermore, we also show that top-ranked opinions of amateur investors can reveal more suitable timing on relatively lower (higher) price levels in the bullish (bearish) market than the opinions of analysts.

5.2 Case Studies

Although we show the differences between professional analysts and amateur investors' reports from readability and wording aspects, it may be still hard to imagine these market participants' reports. This section provides some case studies to discuss what will be written down in both reports and what will only exist in the reports from a particular group.

Figure 2 shows an example of both reports. We highlight each paragraph and adopt the notion of argument mining [31] to mark these paragraphs. Most analysts describe some recent facts, and then provide some analysis (rationales) to support their claims. Some of them may provide suggestions at the end of the report. In amateur investors' posts, they will also describe events, provide some analysis, and make claims. However, there are some differences as follows. First, the amateur investors may use rumors (hearsay) as rationales to support their claims, but the professionals may not use

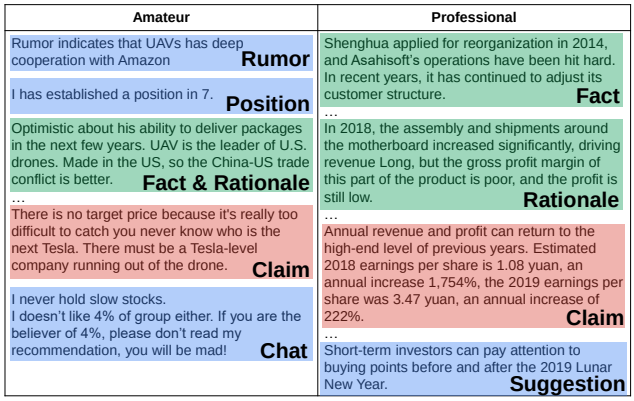

Figure 2: Example of amateur's and professional analyst's reports.

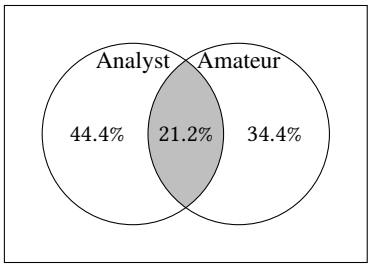

Figure 3: Venn diagram of mentioned stocks.

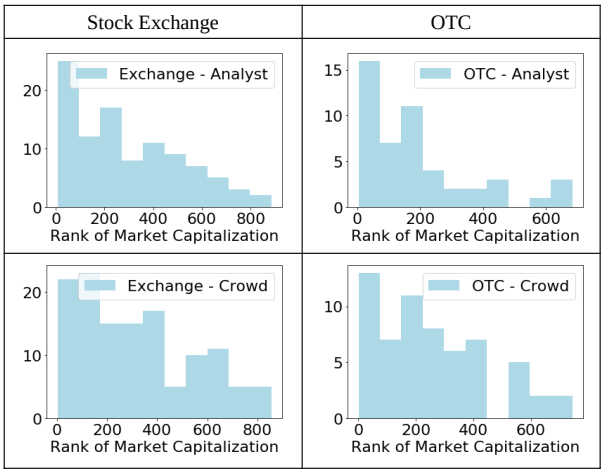

Figure 4: Comparison of the mentioned stocks from market capitalization ranking aspect.

Table 8: Statistics of mentioned stocks.

	Crowd	Analyst
Stock Exchange Market	67.72%	66.89%
Over-the-Counter Market	32.28%	33.11%
Average Market Capitalization (Million)	12,556	12,846

the unverified information. Second, some amateur investors may discuss their position, and few analysts show this information in their reports. Third, some amateur investors may chat with other social media users in their posts. Although it can be sometimes considered as a suggestion, it is chit-chat in most cases.

On the other hand, whether the professionals and amateurs discuss the same stock in the same period is also an interesting question. To answer this question, we compare the analysts' reports (1,029 reports) collected from Bloomberg Terminal from Dec. 2018 to Jun. 2019 with 662 posts on the social media platform during the same period. During this period, 294 stocks were mentioned in analysts' reports, and 249 stocks were mentioned by amateur investors. Only 95 stocks were mentioned in both groups. These statistics show that the amateur could provide additional information to those not analyzed by professional analysts, and vise versa. That is, professionals and amateurs may pay attention to different financial instruments. We also find that one high-quality post published on 2019/06/11 mentions the same stock that is also in the analyst's report on 2019/04/15. However, the sentiments of the post (bearish) and the report (bullish) are different. It evidences the opinions of professionals and amateurs are complementary.

In order to find out what the amateurs pay attention to and what the professionals focus on, we compare the market capitalization of the stocks mentioned by the investors. Figure 4 shows that the crowd tends to mention more small market capitalization stocks in both stock exchange market and over-the-counter market (OTC) than professionals. It provides one possible reason for the complementary phenomena. We further show the details of the mentioned stocks in Table 8. The results indicate that the proportion of the stocks mentioned in both the stock exchange and OTC markets is similar in both groups. Additionally, we also find that the average market capitalization of stocks mentioned by the crowd is about 300 million lower than that of the stocks mentioned by professionals. It echos the findings in Figure 4, again.

Whether professionals and amateurs mention the same target on the same day is also an interesting question. To answer this question, we provide the statistics based on the 95 overlap stocks. Since some stocks may be mentioned several times during the period, we use the first-mentioned date in both groups for analysis. About 54.74% of stocks are mentioned by amateurs earlier. That means it is hard to say whether the amateurs follow the opinions of professionals. It is more likely that the investors in each group analyze stocks based on their own views, and sometimes they may find the same stock's potential. We analyze two cases that amateurs mentioned one-day earlier and two cases that professionals cited one-week earlier, and get the following findings.

- Both amateurs and professionals may release their analyses right after the monthly earnings report or the annual report. Since amateurs can post their analyses anytime, they sometimes release reports earlier than professionals.
- In the posts where amateurs mentioned the same stock one-week later than the professionals, "the quantity bought by

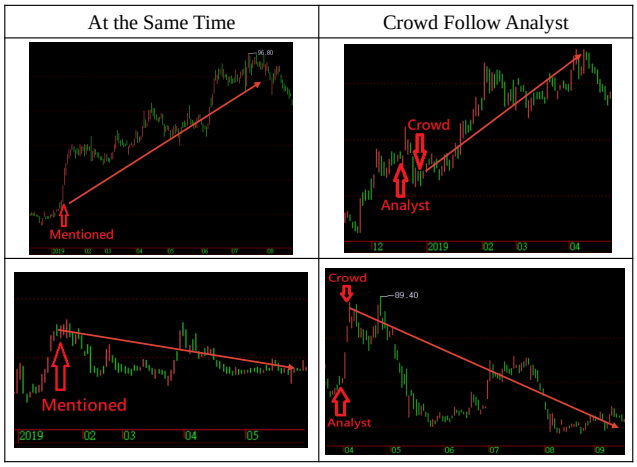

Figure 5: Stock price movement of four cases.

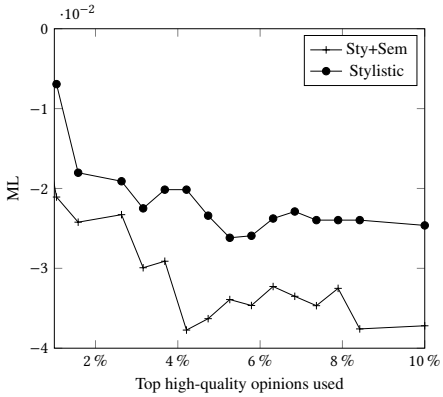

Figure 6: Average ML under different numbers of high-quality opinions. "Sty + Sem" denotes "Stylistic+Semantic".

foreign institutions"[5] is one of the listed reasons. It shows that sometimes the amateurs follow the trading of professionals.

- One amateur sets a higher price target than the professional, and one amateur makes a higher EPS forecasting than the professional. On the other hand, one amateur sets a lower price target than the analyst, and one amateur does not provide the prediction of price or earnings. It may be hard to compare the price targets and the EPS forecasting of the investors in both groups by only four cases. However, in our previous work [8], we show that the price targets from crowd investors are more progressive than that of professional analysts.
- All cases are bullish posts, so that we can analyze how the market reflects the investors' opinions. Since those cases in which crowd investors post earlier than professionals are only one-day early, we assume that both amateur and professional mention at the same time. Figure 5 shows the results of these cases.That remains one of the possible research directions for future works: how the market reacts to the different cases, including "mentioned at the same time", "crowd follow analysts", and "analysts later than crowd". In our previous work [5], we demonstrate one possible approach to test the informativeness of the trading signal or events with the Kolmogorov-Smirnov test.

5.3 Comparison between Different Ranking Methods

As we show in Section 5.1, using semantic information better discriminates expert rationales from crowd rationales because of the varied wordings between these two groups. To answer **(RQ2)**, however, stylistic features are more proper for capturing expert-like opinions. There are two possible reasons explained below. Firstly, compared with the approach using both stylistic and semantic features, using stylistic features only reduces 33.87% of the downside risk when we use the opinions of the first decile. In Figure 6, we compare the average ML under different numbers of high-quality opinions, and provide other evidence to the effect that stylistic features are best for controlling downside risk. Secondly, as we show in Table 6, analysts tend not to use negative words. Due to the wording habits of professional analysts, it may mean that we cannot glean bearish opinions from amateur investors when we adopt semantic features. With stylistic features only, we can remove this restriction.

[5] After the trading hours, the Taiwan Stock Exchange Corporation (TWSE) release the trading volume of foreign and local institutions every day.

Table 9: Aggregation of analysts and top-ranked posts.

	Average MPP	Average ML
Analyst + Sty	**20.55%**	**-5.43%**
Analyst + Sty + Sem	19.43%	-6.67%
Analyst + User Feedback	19.26%	-7.01%

According to Figure 7, we find that only 10.26% of posts are the same under different ranking approaches. Most top-ranked posts of our approach are different from those ranked by user feedback, and over 50% of top-ranked posts mined using stylistic features are different from those mined using both stylistic and semantic features. This suggests that the results mined by different features have different characteristics. Here, we have already compared these approaches from both profitability and risk aspects. We leave analysis from other aspects and the development of ensemble ranking approaches to future work .

Finally, we want to discuss one of the possible real-world applications of the proposed method. The investors can construct a portfolio based on the opinions of both professionals and the crowd. Table 9 shows the results of aggregating analysts and top-ranked posts. We find that if the investors trading based on (1) the first decile opinions sorted out by stylistic-based method and (2) the opinions of professionals, the average MPP and ML are 20.55% and -5.43%, respectively. That means we can reduce the downside risk by lower the potential returns. These results also support that using stylistic features is better than using both stylistic and semantic features, again.

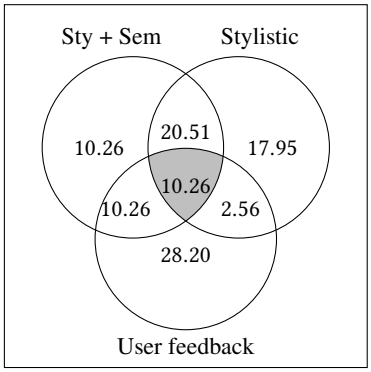

Figure 7: Venn diagram of documents in the first decile under different ranking methods. All numbers are in percentage (%). "Sty + Sem" denotes "Stylistic+semantic".

6 ADVANCED EXPLORATION

As shown in Section 3.2, we separate the rationales from the other descriptions in the analysis reports of both professional analysts and amateur investors by using the subtitles in these reports. That shows our experiments are based on the well-formed documents and leads to another research question: can we detect the rationales of investors from free-formed documents? In order to explore this question, in this section, we propose a pilot dataset, named **I**nvestor's **C**laim-**R**ationale **D**ataset (ICRD)[6], and probe several models to show the performance on free-formed document understanding.

6.1 Tasks in ICRD

There are two tasks in the proposed dataset, including (1) rationale detection and (2) claim-rationale inference. In the rationale detection task, we ask models to detect whether the given sentence is a rationale. For example, "Earnings growth is lagging peers and volatility is significantly higher than processed food ones" is a rationale. How to align the claim and the rationales is also an unexplored task in the financial narrative. Thus, we propose the second task, named claim-rationale inference. Given a claim, we need to find the rationales that support the given claim. In this way, we can align the claim with the rationales.

6.2 Construction of ICRD

Two experts working in the financial industry are involved in the annotating process. Given a sentence from the analysis report, they are asked to label whether the given sentence is the "claim" of the investor. If the given sentence is the claim of the investor, they further need to separate other sentences in the same paragraph into two classes, including (1) "rationale" of the selected claim or (2) "not rationales" of the selected claim. The third annotator is involved in checking for those instances getting different labels. Finally, we get 21,444 claim-rationale pairs in this dataset. We use 70% of instances as the training set, 10% of them as the development set, and 20% of them as the test set.

[6]ICRD: http://nlg.csie.ntu.edu.tw/nlpresource/ICRD/

Table 10: Macro-F1 scores of the experiments on ICRD. Inference denotes the claim-rationale inference task.

Model	Claim Detection	Rationale Detection	Inference
CNN	76.15	55.25	53.75
BiGRU	77.97	48.62	54.74
CapsNet	77.93	52.47	51.97
BERT	**79.86**	**57.69**	**56.96**

6.3 Experiment on ICRD

In addition to experimenting on ICRD, we also report the results of the claim detection task proposed in our previous work [7]. In the claim detection task, we aim at detecting whether the given sentence contains a subjective opinion from the investor. For example, "We upgrade our price target to 2,000" is a claim of an investor. We adopt CNN, BiGRU, capsule network (CapsNet) [29], and BERT to explore the proposed tasks. Because we consider all tasks as a classification task, we use the macro-F1 score as the metric for evaluation. Note that, in the claim-rationale inference task, we ask models to determine whether the given sentence is the rationale supporting the given claim. Table 10 shows the experimental results. We find that detecting the claims is easier than detecting the rationales. The reason may be that most investors use the same words to present their claims such as "estimate", "price target", "upgrade", and "downgrade".

The experimental results also show the difficulty of the claim-rationale inference task. Based on the results of claim detection and rationale detection tasks, we can get a clue for this phenomenon. It may be caused by the similarity of the narrative style of the rationales and that of other sentences that describe the facts. For example, "BRF shares are down by 39% YTD" is not a rationale related to any claim in the report, but "PPOP was only down by 2% yoy" is a rationale of the investor for supporting the estimation on the stock price. Both sentences are similar but have different meanings. This makes the tasks related to rationales detection more difficult.

In this section, we probe the rationale detection issue on free-formed documents. With the proposed dataset and the experiments with several models, we provide the first exploration of this topic and show the difficulty of the proposed tasks. However, based on our findings in Table 4, we can get more profit and take less risk by using the results of evaluating the rationales of the investors. Further exploration of detecting the rationales is still needed. We will release the annotations of the proposed dataset under the CC BY-NC-SA 4.0 license.

7 FUTURE RESEARCH DIRECTIONS

In the past, many works [4, 11, 24] focus on the sentiment only in financial opinion mining, and few discuss the rationales and the quality of the investors' analysis. In this paper, we show the importance of evaluating the rationales of investors. The experimental results support that the proposed direction is promising. The concept of these explorations can also be considered as the investigation from opinion mining in finance to financial argument mining. To facilitate the development of this direction, we present the concept of argument mining in financial narratives in this section as the suggestions for future works.


J.P.Morgan

Michaels

Overweight

MIK, MIK US

Price (16 Sep 20): $10.18

Price Target (Dec-20): $16.00

We expect the following: (1) Long-term targets of LSD SSS, MSD EBIT growth and HSD EPS growth driven by MIK's ongoing retail 101, omni-channel, and makers/Pro initiatives to drive topline/share (see bullet below) with the opportunity to improve margins through labor efficiency, merchandising rigor, inventory flow disciplines, cost leverage, and sourcing/private label expansion. **(2)** Capital/investment spending to remain relatively consistent with history given modest new store growth and a highly manageable omni-channel investment cycle (i.e., no need for a big supply chain or tech stack buildout); MIK targeted 2.5-3.0% of sales for capex on its

Figure 8: Professional analyst's arguments.

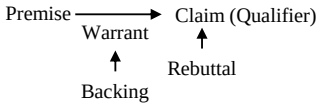

Figure 9: Toulmin's argumentative model.

7.1 Elementary Argumentative Units

Argument mining is one of the hot topics in recent years [3, 21, 26]. Different from opinion mining and sentiment analysis, which focus on predicting the main claim (positive/negative) only, argument mining aims to analyze the arguments (claims and premises) supporting the main claim and further evaluate the persuasiveness and the rationality of a document.

As one of the fundamental rhetorical modes, argumentation is frequently applied in financial narratives. Taking Figure 8 as an example, opinion mining researches pay the attention to predicting the market sentiment (overweight/neutral/underweight) based on the narrative in the report, but do not take the reasons and the persuasiveness or rationality of these reasons into account. Imagining that if there are two reports, one provides overweight rating to the stock of The Michaels Companies, Inc. and the other one gives underweight rating to this stock, should we conclude that currently investors' market sentiment is neutral to this stock? Clearly, the answer to this question should be no. Thus, how to compare the arguments in both reports and evaluate their rationality is the next step that we should focus on. In Section 6, we explore the first step to separate the free-form narrative into claims and rationales (premises). As the start point for entry financial argument mining, in this section, we introduce the elementary argumentative units of investor's opinion in Figure 8 based on the notion of Toulmin's argumentative model [32] shown in Figure 9.

Claim and premise are two basic units in an argumentation. Claim is the subjective view of the investor, and premise is the objective facts used to support the claim. In Figure 8, the underlined sentences are the claims of the analyst, and the other sentences in the same point are the premise of the claim. That is, premise can also be called rationale. Warrant is the background knowledge that makes the investors infer the claim based on the premise, and backing is used for supporting the warrant. For example, in the first point in Figure 8, the analyst infers a claim ("EPS growth") based on the premise ("improve margins through labor efficiency"). The warrant is that we can make more products in the same work hours with the improvement of labor efficiency, and it will also lead to the growth of income. In this case, the backing is the common sense in accountancy. In most real-world scenarios, the warrant and backing are implicit information in the argumentation. Normally, it is not written down in the documents.

In argumentative models, qualifier stands for the strength of the claim, which can be the rationality of the inference or the confidence of the investors. In Figure 8, we can take the price target as a proxy for the confidence of the analyst. That is, the analyst concludes that the stock price will rise from 10.18 to 16, and the difference between the close price and the price target, i.e., 48.18%, could be the qualifier of this report. Finally, rebuttal stands for the counterarguments that defeat the claim. Because the report in Figure 8 is the opinion of one analyst, there do not exist rebuttal cases. We will explain rebuttal in detail in Section 7.3.

7.2 Argumentation Structure in an Opinion

After extracting each argumentative unit, we need to link these units and make inferences based on the extracted information. For example, in Section 6, we infer whether the given description can support the given claim. Figure 10 shows the argumentation structure of the analysis report in Figure 8. In this report, the main claim (MC) on the stock of Michaels is overweight. The analyst makes six claims (c) to support the main claims, and each claim is supported by a different number of premises (p). The structure from p_1 to the MC is called a sequential structure, where w denotes the rationality of the premise to the claim and q stands for qualifier. The structure of (p_2, p_3, c_2, c_3) is named linked argument, where p_2 supports c_2 and c_2 is also supported by c_3 with p_3. Some claims like c_4 may not be supported by any premises. The structure of (p_4, c_5, c_6) is a divergent argument, where two claims are supported by the same premise. The full argumentation structure is a hybrid structure. Previous works show that encoding the argumentation structure into models is useful for evaluating the quality of persuasive essays [34] and the persuasion of online debates [22]. However, few studies adopt the same idea for analyzing the investor's opinion. In Figure 10, we not only provide an example of forming the investor's opinion in an argumentation structure, but also indicate that evaluating each pair's rationality and giving weights for the edge can help us better understand the financial narratives. With the rationality scores, the argumentation structure becomes a directed weighted graph. This kind of anatomy is much closer to an investor's behavior when reading a report.

7.3 Argumentation Structure of Opinions

In financial market, investors debate on the price movement all the time. Figure 11 shows an example of the argumentation structure of the opinions, comprising a discussion on an online forum. The original post makes a claim on the TSM's price and provides several premises from different aspects. The first reply (R1), which agrees with the original post, can be considered as supporting the main claim of the original post. The second reply (R2) supports one of claims of the original post. The third and the fourth replies, R3 and R4, attack the main claim of the original post from different aspects. In this case, R3 and R4 are the rebuttals of the claim in original post.

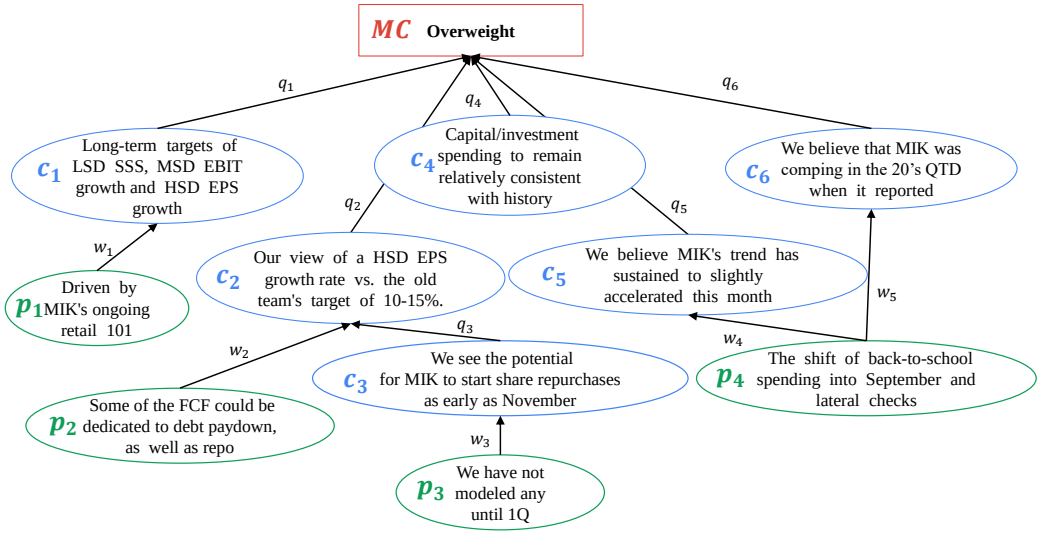

Figure 10: Argumentation structure of the report in Figure 8.

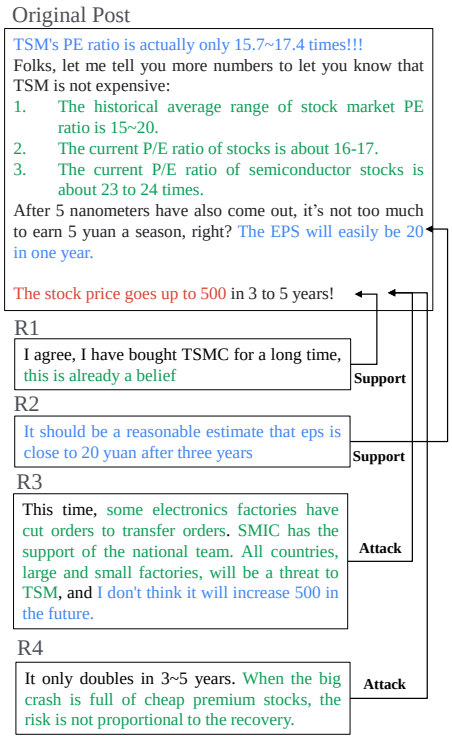

Figure 11: Argumentation structure among opinions.

On an online debate platform, debaters discuss the given topic for several rounds, which is similar to the online financial forum discussions. Investors discuss the possible price movement direction for several rounds from different aspects. In this way, we can adopt the concept of support/attack from argument mining to evaluate the rationale or persuasiveness of the original post. We can further construct a larger argumentation structure, where all arguments of the investors are connected with the edges denoting bullish/bearish stance toward certain financial instrument. By comparing the rationales from the investors stand in both stances, we could not only link the opinions from different investors and different documents to a graph, but also provide an explanation on the decision process.

7.4 Opinion Quality Evaluation

Because the arguments and the main claim are closely interconnected and inseparable, to determine the weights on the edges (w and q) in argumentation structure is the topic that we need to explore in the future. In most arguments, the qualifier is an implicit term and is hard to be inferred. In financial narratives, some information can be used as the qualifier. For example, the price target may be a good proxy for evaluating qualifier. In Figure 8, the professional analyst sets the price target at \$16. In the original post in Figure 11, the amateur investor expects the price will go up to \$500. These instances show that investors make a claim with some estimations. By comparing the estimation with the market information, we can get the qualifier of the investor's claim. However, this kind of information is related to the confidence of the investor instead of the quality of the claim. For example, the close price of TMC at the post date of the original post in Figure 11 is 280. That means the amateur investor sets a progressive price target, and the qualifier of this post could be $\frac{500-280}{280} = 78.57\%$. Should we evaluate the trustworthy or the quality of the claim based on the qualifier? In our opinion, the answer may be no. Evaluating the premises supporting the claim can be explored to select trustworthy or high-quality opinions. That is, estimating w in Figure 10 is more important for finding good claims. In the future, the argumentation structure in financial narratives could be encoded to neural network models. Based on the experience of other works in argument mining [22, 34], we expect that downstream tasks can be improved by the fine-grained analysis of the investor's opinions.

In this paper, we address the problem that evaluating the rationales of a given investor's analysis. In addition to evaluating the rationales of a single opinion, constructing the argumentation structures based on multiple opinions discussing the same financial instruments can also help us understand other investors' views. The notions of "support" and "attack" in Figure 11 are also necessary weights that we need to analyze in the future. The support posts could increase the trustworthy of the target post, and may provide additional premises for the target post. Because most investors only write the analysis to support their main claim, we can seldom see the discussion from both stances in a single post. Thus, the attack posts are very important because they could provide opinions from opposite viewpoints. Furthermore, the rationality of the support and attack posts may also influence the effectiveness of these posts toward the target post. Although the proposed directions are intuitive actions for humans when reading an investor's report, few works explore these directions in financial narratives to the best of our knowledge. We believe that our community will become much closer to human-level language understanding in the financial domain with the proposed ideas of financial argument mining.

8 CONCLUSION

We present an important task—mining high-quality opinions—for research focusing on extracting or using opinions from user-generated textual data. We further propose a novel approach to infer opinion quality by how "expert-like" the rationale supporting the opinion is. Experimental results show the effectiveness of our approach and the usefulness of using top-ranked opinions. We further show that top-ranked crowd opinions mined by our approach are comparable with the opinions of professional analysts in terms of controlling downside risk. We also provide an in-depth discussion of expert writing styles and wording. The future research directions are presented and explored with the proposed pilot dataset, ICRD.

In the future, we plan to extend our approach to different application scenarios, such as (1) finding online reviews worth consulting by comparing the rationales of professional commentators with the crowd and (2) evaluating the credibility of online articles by comparing the news articles of professional journalists.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 109-2634-F-002-040, and MOST 109-2634-F-002-034.

REFERENCES

[1] Matthew Backus, Thomas Blake, Jett Pettus, and Steven Tadelis. 2020. *Communication and bargaining breakdown: An empirical analysis*. Technical Report. National Bureau of Economic Research.

[2] Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. You Write like You Eat: Stylistic Variation as a Predictor of Social Stratification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2583–2593. https://doi.org/10.18653/v1/P19-1246

[3] Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis.. In *IJCAI*, Vol. 18. 5427–5433.

[4] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. NTUSD-Fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*.

[5] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Crowd View: Converting Investors' Opinions into Indicators. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6500–6502. https://doi.org/10.24963/ijcai.2019/936

[6] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral Attachment with Auxiliary Tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1161–1164. https://doi.org/10.1145/3331184.3331361

[7] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NumClaim: Investor's Fine-Grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 1973–1976. https://doi.org/10.1145/3340531.3412100

[8] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 136–143.

[9] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, 740–750. https://doi.org/10.3115/v1/D14-1082

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078* (2014).

[11] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. Association for Computational Linguistics (ACL).

[12] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Magdalini Eirinaki, Shamita Pisal, and Japinder Singh. 2012. Feature-based opinion mining and ranking. *J. Comput. System Sci.* 78, 4 (2012), 1175–1184.

[14] Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. 2019. #YouToo? Detection of Personal Recollections of Sexual Harassment on Social Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2527–2537. https://doi.org/10.18653/v1/P19-1241

[15] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1837–1848. https://www.aclweb.org/anthology/C18-1156

[16] Jyun-Yu Jiang, Xue Sun, Wei Wang, and Sean Young. 2019. Enhancing Air Quality Prediction with Social Media and Natural Language Processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2627–2632. https://doi.org/10.18653/v1/P19-1251

[17] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[19] Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[20] Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with Convolution Units to Predict Stance and Rumor Veracity in Social Media Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5047–5058. https://doi.org/10.18653/v1/P19-1498

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

[22] 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)*. Association for Computational Linguistics, Online, 8905–8912. https://www.aclweb.org/anthology/2020.emnlp-main.716

[23] Quanzhi Li and Sameena Shah. 2017. Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In *CoNLL*. Association for Computational Linguistics, Vancouver, Canada, 301–310. https://doi.org/10.

18653/v1/K17-1031

[24] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Zihan Liu, Yan Xu, Cong Gao, and Pascale Fung. 2019. Learning to Learn Sales Prediction with Social Media Sentiment. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 47–53. https://www.aclweb.org/anthology/W19-5508

[25] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Melbourne, Australia, 698–708. https://doi.org/10.18653/v1/P18-1065

[26] Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7, 1 (2013), 1–31.

[27] Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically Identifying Complaints in Social Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5008–5019. https://doi.org/10.18653/v1/P19-1495

[28] Masoud Rouhizadeh, Kokil Jaidka, Laura Smith, H. Andrew Schwartz, Anneke Buffone, and Lyle Ungar. 2018. Identifying Locus of Control in Social Media Language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1146–1152. https://doi.org/10.18653/v1/D18-1145

[29] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *NeurIPS*. 3856–3866.

[30] Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Melbourne, Australia, 2104–2115. https://doi.org/10.18653/v1/P18-1196

[31] Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies* 11, 2 (2018), 1–191.

[32] Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

[33] Hou-Chiang Tseng, Berlin Chen, Tao-Hsing Chang, and Yao-Ting Sung. 2019. Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering* 25, 3 (2019), 331–361. https://doi.org/10.1017/S1351324919000093

[34] 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*. 1680–1691.

[35] Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. Recommendation in Internet Forums and Blogs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 257–265. https://www.aclweb.org/anthology/P10-1027

[36] Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Berlin, Germany, 195–200. https://doi.org/10.18653/v1/P16-2032

[37] 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*. Association for Computational Linguistics, Melbourne, Australia, 1970–1979. https://doi.org/10.18653/v1/P18-1183

[38] Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018. Multi-Task Learning Framework for Mining Crowd Intelligence towards Clinical Treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 271–277. https://doi.org/10.18653/v1/N18-2044

[39] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2968–2978. https://doi.org/10.18653/v1/D17-1322

[40] Annie Ying and Pablo Duboue. 2019. Rationale Classification for Educational Trading Platforms. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 14–20. https://www.aclweb.org/anthology/W19-5503

[41] Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network. In *WWW*. ACM, 3448–3454.