Financial Argument Quality Assessment in Earnings Conference Calls

Alaa Alhamzeh(✉)

University of Passau, Passau, Germany
Alaa.Alhamzeh@uni-passau.de

Abstract. Arguments are ubiquitous. Yet, the definition of what is a good argument depends on the goal, and the settings. Although various business communication studies confirm the crucial role of argumentation, no work has shaped financial argument quality in a way that is concise enough for a practical application.

In this paper, we aim to close this research gap by modeling the quality of managers' arguments, during the Q&A sessions of earnings conference calls. To this end, we propose various quality dimensions at both levels of argument and argument units. Our quality model establishes a well-considered link between, on the one hand, insights as they are expressed in financial text analysis literature, and, on the other hand, insights derived from empirical quality descriptions as provided by argumentation discourse linguistics and computational models.

We further conducted the related annotation study and produced *FinArgQuality*, the first financial dataset annotated with argument quality. This corpus composes of 14,146 sentences in a total of 80 earnings calls transcripts.

Our proposed quality assessment dimensions, and final annotated corpus are publicly available, and can serve as strong baselines for future work in both FinNLP and computational argumentation disciplines. We further discuss some potential optimization goals and financial applications of this data, and highlight future directions.

Keywords: Argument quality · FinNLP · Earnings Conference Calls · FinArgQuality dataset

1 Motivation

The rise of data and the development of machine learning have arrived at the foundation of the financial technology (FinTech) domain. This interdisciplinary field aims at supporting financial services with digital innovations and technology-enabled business models [1]. However, given that about 80% of today's data is unstructured information which is composed mainly of textual data, some argue that Natural Language Processing (NLP) is the most important research nowadays. That is because we need to automatically leverage this data and to process it in different frameworks, which, in financial aspect, is called Financial Natural Language Processing (FinNLP).

C. Strauss et al. (Eds.): DEXA 2023, LNCS 14147, pp. 65–81, 2023.
https://doi.org/10.1007/978-3-031-39821-6_5

Among various financial textual sources, Earnings Conference Calls (ECCs) are one of the most considerable information players in the stock market [2].

An earnings call is a quarterly organized event where public traded companies report their last quarter performance and give guidelines about the next one. The company management often discuss and detail key points, such as growth, risks, buybacks, and dividends. Explicitly, an earnings call consists of two sections: a presentation held by the company, followed by a Questions & Answers (Q&A) session where company representatives[1] answer the questions of professional analysts and other market participants. The analysts are later expected to announce their opinions about this stock in a sort of recommendation (in a scale 1–5), expected price target and sometimes a detailed report. In fact, many studies show that analyst's discussions during the question-answering session to be the most informative and impacting part on the market [3–5]. Therefore, we focus in our study on this particular section of the call.

During this session, the management team may provide additional context and information on the company's financial results and future outlook, which can help analysts better understand the company's performance and make more informed recommendations. For example, if a company reports weaker-than-expected earnings, but the management team explains that the results were affected by one-time events that are not expected to reoccur, analysts may be more likely to maintain or even upgrade their recommendations. Contrarily, if a company reports weaker-than-expected earnings, and the management team provides no clear explanation for the results, analysts may be more likely to downgrade their recommendations towards this company.

In fact, the automatic analysis of earnings calls is valuable for different financial services and applications (e.g., financial risk prediction [6,7], modeling of analysts' decision-making [2]). However, these calls are still an under-resourced text genre in computational argumentation, despite the fact that various business communication studies proved the important role of argumentation in ECCs (e.g., [8,9]).

In our previous work [10], we covered the argument structure in the managers' speech during the Q&A sessions of four top tech companies *(Apple, Facebook, Amazon, and Microsoft)* for the period of five years 2015–2019, resulting in 80 transcripts.

In this work, we extend on it, to study further the quality of managers' arguments during these public calls. In other words, we want to answer the following research question: How to handle the quality of company executives' arguments, while establishing a well-considered link between, on the one hand, insights as they are expressed in financial text analysis literature, and, on the other hand, insights derived from empirical quality descriptions as provided by argumentation discourse linguistics and computational models?

We tackle this research gap by conducting a comprehensive synthesis on earnings calls and Computational Argument Quality (CAQ) state of the art. Investigating on the same *FinArg* corpus, we have introduced in [10], our contributions

[1] Mainly chief executive officer and chief financial officer.

in this paper are two-fold. First, we develop a scoring model for argument quality in ECCs, that covers both argument units and overall argument. Second, we conduct the related annotation study, and contribute to the research community with *FinArgQuality*: the first financial corpus annotated with argument quality scores.

This paper is organized as follows: Sect. 2 covers a conceptual background of argument quality, financial text analysis, aligned with most related works to our research. In Sect. 3, we define and illustrate our argument quality dimensions with detailed examples and explore our rating rubrics. We further report our annotation study, inter-annotator agreement, size, and statistics of our final corpus - *FinArgQuality* in Sect. 4. We discuss our findings, the potentials of this data, and conclude future directions in Sect. 5.

2 Related Work

Our work is closely related to the following two lines of research:

2.1 Computational Argument Quality Assessment

An argument is defined as the justification made to reach a conclusion on a controversial topic. Thus, the simplest argument composes of one claim and one premise supporting it. Argument Quality is the assessment of its attributes, strength, and persuasiveness. Delving into the rich realm of argumentation theories, various quality proposals have been introduced. To the best of our knowledge, the computational argumentation literature reported only one study that comprehensively survey the argument quality assessment theories and proposals by Waschsmuth et al. [11]. By that, they derived a taxonomy of 15 dimensions covering the logical (e.g., level of support), rhetorical (e.g., persuasiveness) and dialectical (e.g., relevance) aspects of an argument.

Despite the philosophical background of argumentation and argument quality, researchers in computational argumentation looked for practical, yet considerable definitions of argument quality. They further faced this problem with different methodologies of assessment. An overview of the literature approaches entails the following categories of treatment: First, *Point-wise versus Pair-wise Rating*, meaning, either an absolute rating of the argument (e.g., ranking the strength of a student essay [12]), or a relative rating of it in comparison with another argument (e.g., which argument is more convincing by [13]). Second, with respect to the *level of granularity*, we can distinguish methods that estimate the quality of an overall argument (e.g., [14]) versus, the quality of its particular components (e.g., [15]). Furthermore, some scholars explored the interaction dynamics into debate context. For instance, [16] tried to define the wining argument on the Reddit platform using the interaction patterns. Third, regarding *the method of assessment*, the literature reported mostly direct classification (regression) models (e.g., [17,18]), with some indirect attempts. For instance, [19] investigated on a set of linguistic features that reflect the argument quality

instead of considering the original text. Similarly, Gurcke et al. [20] aimed at assessing the sufficiency of arguments through conclusion generation. However, not surprisingly, direct methods outperform their peers. This discussion should give you a bird's-eye view on the diversity of computational argument quality field.

Furthermore, while many studies treat the argument in a holistic manner, Walton [21] argues, *"if the concept of an argument is defined in terms of the premises in it (providing grounds or reasons for accepting the conclusion), then we have to ask what "grounds" or "reasons" are, other than being good or reasonable arguments"*. We also follow this vision in our argument quality dimensions. Thus, we distinguish further the types of argumentative units (i.e., premises and claim). We provide further discussions all across our quality dimensions.

2.2 Text Quality in Finance and Business Communication

The analysis of available textual data has always been a topic of interest for many researchers in the financial domain. However, the end target could be widely different. For example, while [22,23] evaluated the forecasting skills of investors, [24] analyzed the managers' speech with the goal of predicting the financial risk, and [25] aimed at making a future price prediction out of detected events on news and social media. We present in the following some related work that is directly linked to our proposed quality metrics:

Zong et al. [23] used the Linguistic Inquiry and Word Count (LIWC) lexicon [26] to detect the temporal orientation of a forecaster's justifications. They found that good forecasters tend to focus more on past rather than future events. Therefore, we build on that, and we extend to more fine-grained assessment of the past level in our *temporal_history* attribute.

Besides, as we have aforesaid, various business communication studies proved the important role of argumentation in earnings conference calls. Among others, Rocci et al. [27] differentiate evidential type presented in different sections of an earnings calls to be: "common knowledge, direct, epistemic possibility, generic indirect, inference, report, and subjective". In their empirical study, they found that the subjective type to be the most frequent in the answers of company executives. Hence, we consider studying the *subjectivity* of an argument as one of our quality metrics, since we want to highlight the objective arguments.

Notably, different financial studies focus on the statement specificity as a major factor of its quality. Text "uncertainty" [23] and "hedging" [2] are only indicators of "the lack of commitment to the content of the speech" [28]. This is logical, since the qualitative analysis of a financial text cannot be separated from its quantitative property. Therefore, we also concentrate on the argument *specificity*, but further from two angles: the specificity of the answer in relation to the asked question, and the specificity of the premises and claims through identifying their particular types.

3 Argument Quality Dimensions

Given that the criteria of what is a good argument depends on the goal orientation [19,29], we define our quality attributes in collaboration with experts from the Chair of *Financial Data Analytics* at Faculty of Business, Economics, and Information Systems - University of Passau[2].

The CAQ literature reported different guidelines in terms of the annotation scale. For instance, Stab et al. [17] reported 681 (66.2%) sufficient to 348 (33.8%) insufficient arguments in their student essays corpus. Likewise, in the corpus of Persing and Ng [12] annotated with the strength attribute of 1000 student essays, they used a scale of 1.0 to 4.0 with 0.5 increments, giving a total of seven values. Among all essays, 372 are categorized as class 3.0, whereas only 2 are categorized as class 1.0; 21 with class 1.5 and merely 15 belong to class 4.0. Therefore, to avoid such a high data imbalance and to make more fair fine-grained judgment rather than binary decision, we suggest our annotation guidelines with respect to a 3-point scale of assessment, except for *objectivity* which remains binary, given its nature, and the *temporal-history* of an argument, to provide more gradual indicators. In addition, we suggest the argument quality on the unit level (claim and premise types) to be categorical rather than numerical.

In summary, our rating follows the point-wise approach, and looks at each argument from two levels:

3.1 At the Level of Argument

A holistic assessment of an argument quality is the most used approach in the literature. We present in the following the quality metrics we define at the granularity of the overall argument. In other words, considering the argument claim and premises as well as the relations between them.

- **Strong** Persing et al. [12] labeled the strength of a student essay using a scale 1.0 to 4.0. On the other hand, [30] inspected the strength of only the premise component. They defined it by "how well a single statement is contributing to persuasiveness" on a scale 1–6. Inspired by these studies, we define the strength of an argument by two factors: *how many and what type of premises are backing its claim?* For example, an argument with a statistical premise is supposed to be stronger than an argument with a hypothetical premise. Furthermore, Table 1 represents the rubrics for rating the argument strength.
- **Specific** Carlile et al. [30] studied the specificity of every single argumentative statement in a student essay (i.e., premise, claim, major-claim). They score it on a scale of 1 to 5 based on how detailed the statement is. The main source of tolerant and inexact language is using hedging expressions. Prokofieva et al. [28] defined some general guidelines for recognizing hedge expressions in English. Hedges can appear in forms like: "I think", "it is sort of", "probably", etc. In our particular case, we study the arguments presented by company managers to answer analysts' questions. Therefore, it was important for us to

[2] https://www.wiwi.uni-passau.de/en/financial-data-analytics.

declare the specificity in a relation to the question itself. Hence, we rate the argument specificity on a 0–2 Likert scale, as illustrated in Table 1.

- **Persuasive** The persuasiveness is the most subjective attribute to judge. Yet, it is still taken into account by many other studies. This could be due to the fact that, we have a more holistic feedback from the annotator about all argument elements, and their coordination. In addition, we can use these annotations to analyze the relations with other argument attributes (i.e., what makes a persuasive argument). Table 1 displays also our hints to label persuasiveness across arguments.

Table 1. Quality dimensions at the argument level.

Attribute	Definition	Score
Strong	How well the statement contributes to persuasiveness, considering the count and types of supporting premises?	• Strong-0: A poor, not supported argument (e.g., the claim is supported by only one premise that is doubtful) • Strong-1: A decent, fairly clear argument. The argument has at least two premises that authorize its standpoint • Strong-2: A clear and well-defended argument, supported by concrete and powerful premises
Specific	How well the statement is precise and answers directly the question?	• Specific-0: The argument is not related to the question (e.g., blaming the market, mentioning competitors) • Specific-1: The statement partially answers the question, but still implies some hedging • Specific-2: The argument is concrete and directly related to the question
Persuasive	From the annotator view, to what extent is the argument convincing?	• Persuasive-0: The argument is not easily understandable, the speaker may state some description, incident, value but does not explain why it's important. It may then persuade only listeners who are already inclined to agree with it • Persuasive-1: The argument provides acceptable reasoning, may still contain some defects that decrease its ability of convincing. Hence, it would persuade some listeners • Persuasive-2: A clear, well-structured argument that would persuade most listeners. The speaker stated precise and sound premises that remove doubts of the listener
Objective	Is the argument based on facts rather than feelings or opinions?	• Objective-0: A subjective or biased argument based on particular views and opinions • Objective-1: A logical argument supported by verifiable evidences
Temporal-history	Does the argument include any time indicator? In case of many, choose the most recent one	• Temporal-3: during this quarter • Temporal-2: up to two quarters • Temporal-1: half to one year • Temporal-0: more than one year • Temporal-1: not mentioned (if there is no explicit time indicator, choose this value, even if you think that it could be concluded from the context)

- **Objective** Being objective, is very essential from the market perspective. Arguing by opinions and particular views has less impact on investors than arguing with objective information and reached earnings. Hence, we binary classify the argument to objective or subjective based on the question: is the argument based on facts rather than feelings or opinions?.
- **Temporal-history** The temporal information assessment, composes a special phenomenon in financial opinions. Studying the time associated with given information, and estimating its impact period, are important research questions to the stock market [22,31,32]. On the other hand, in a business communication study, Crawford et al. [33] analyzed the persuasion language in economic "Crisis Corpus" in comparison to economic "Recovery Corpus". They found that executives tend to emphasize progress and future expectations in the crisis corpus, while they report achievements in their recovery time period. This is similar to the findings of [23] we have aforementioned, that providing past information reflects better forecasts. Hence, we ignore future expressions and rather weight the temporal spans of text that represents a real value for finance, by recognizing five degrees of temporal-history as shown in Table 1.

3.2 At the Level of Argument Unit

Most argument models include one type of premise. However, we can easily distinguish different types of premises in everyday discourse [34]. For example, a premise may provide empirical evidence, a fact, or a justification why the reasoning of an argument is correct. Similarly, this applies to claims.

Despite the fact, that knowing the types of the argument claim or premise(s) can give us a clear estimation about its quality, the literature reports very rare attempts towards this research direction. Moreover, the annotation of those types could be more objective and less biased itself than scoring the whole argument towards one attribute (e.g., strength, clarity, etc.). Hence, we elaborated part of the data with one of our annotators and suggest the following pragmatic types of premises and claims, as shown in Fig. 1.

Types of Claims
Clarile et al. [30] distinguish three types of claims: Fact, Value (something is good or bad), and Policy. Their study shows that fact claims seem to be the most frequent in their corpus of student essays. We distinguish the following types of a claim:

- **Fact** The earning conference call, is the event where a company shares private information with the public. Therefore, some managers' claims tend to be facts, that still need to be accepted by supporting evidences.

> Example: *"..When it comes to our Commercial Licensing and our servers, it's the same trend, which is the big shift that's happening is our enterprise and datacenter products, being Windows Server, Systems Centers, SQL Server, are more competitive..."*

- **Value** Considering our kind of data (earnings calls), when claiming some information that reflects quantities and reports measures, the claim is classified as a numerical value.

> Example: *" Secondly to provide a bit more color, sales of the Watch did exceed our expectations and they did so despite supply still trailing demand at the end of the quarter."*

- **Opinion** We identify this type of claims, for all statements that reflect the company vision and its executives' standpoints. Few terms introducing an opinion are like: *we're very happy, I think*. In fact, this type of claim is very common, especially while expressing the company future hopes [33].

> Example: *"...And so we are incredibly optimistic about what we've seen so far."*

- **Policy** This kind of claim is used to express a plan of action, or existent rules.

> Example: *"And so as you know, we don't make long term forecasts on here."*

- **Reformulated** During our pilot annotation, we observed a common pattern of repeating the same claim with some reformulation, mainly at the end of the answer. Hence, we define the Reformulated claim type, which could be justified by the oral argumentation nature of our data.
  According to [35], reformulation or restatement is a rephrase of the evaluative expression without adding any significant information, where the goal is to make certain that the evaluation is clear and unambiguous. Some indicators to reformulations are: *in other words, that is to say, rather*. In our data, the reformulated claim is mostly the shorter one of the two claims. We ask the annotators not to link this claim to any premises (i.e., not to consider it as a new argument).

> Example: *"...And I think when you take those two things, along with what Satya said, being able to balance disciplined focus and execution for us, I think we feel very good about the progress we've made."*

- **Other**
  This label is selected when no particular claim type is recognized.

Types of Premises. Similarly to claims forms, and motivated by the works of [30,36], we set the following premise types:

- **Fact** This unit provides evidence by stating a known truth, a testimony, or reporting something that happened.

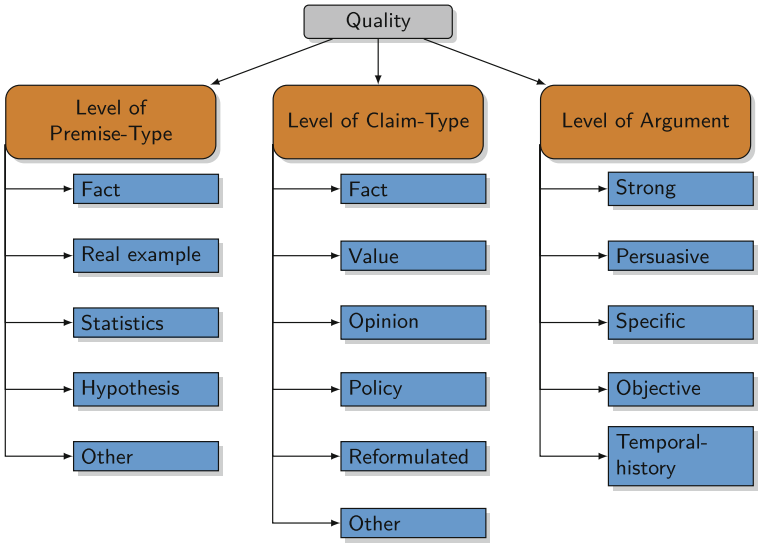

Fig. 1. Our quality dimensions at the levels of argument and argument units

> Example: *"And then at the same time, we're bringing more and more advertisers into the system and that's giving us a better selection of the ads that we can serve to the people using Facebook, and that, again, improves the quality and the relevance."*

- **Real Example** Sharing out a comparable experience, a specific event, or similar, is a common strategy in spoken language and in argumentation in general [37].

> Example: *"I also look at the first time iPhone buyers and we're still seeing very, very large numbers in the countries that you would want to see those in, like China and Russia and Brazil and so forth."*

- **Statistics** This type of premise is decisive in any argumentative discussion. Definitely, it is very common and powerful in earnings calls.

> Example: *"So we ended the year last year with 109 fulfillment centers around the world and 19 U.S. sort centers..."*

  This example also implies that the automatic understanding of numerical data is more complicated in this genre of text [31].
- **Hypothesis** Besides probative deductions, hypothetical, and assumption evidences can be used. However, this type of text seems not to be frequent in our data.

> Example: *And if this works as planned, it can be big.*

- **Other** This unit is supporting the final conclusion, but none of the previous evidence characteristics applies to it.

> Example: *"...These numbers are unbelievable and they're done in an environment where it's not the best of conditions..."*

We assume that those fine-grained types of argumentative units, should provide a clear and concrete reflection of the argument quality. In addition, recognizing the argument ground basis of reasoning is inline with analyzing the argumentation scheme [38].

4 Data Creation

4.1 Annotation Study

We downloaded our data using a paid subscription to the Financial Modeling Prep API[3]. We used Label Studio[4] as a visualized annotation tool. Our annotated data concerns the quarterly earnings calls of four companies: Amazon, Apple, Microsoft, and Facebook for the period of 2015–2019. For each transcript, we determine, using a Python script, the list of all the speakers (Analyst, Representative, or an Operator). We then split each transcript into different documents. Each *document* contains one or two questions asked by a *single analyst*, along with the corresponding response(s) by the company's managers.

We elaborate with one of our annotators to define the annotation guidelines as a first step. Later on, three other annotators have joined the annotation process. All of them have a significant level of language. One of them is an international economics master student, whereas the others are all computer-science students. Therefore, our choice of companies was more tech-oriented, where industry jargon and different products are known to all. Moreover, to let annotators gain insight into the company's performance over the years, we assign one company for each of our four annotators to do all its quarters' annotations. We started by training sessions and discussions with the annotators. Based on their feedback, we were able to refine the guidelines and clarify ambiguous situations. Thereafter, a division of the data was done to 20% to be double annotated for inter-annotator agreement calculations, and 80% individual annotations. We call the output corpus: *FinArgQuality* and it is publicly available to foster future research[5].

4.2 Inter Annotator Agreement

To calculate the inter annotator agreement, we define about 20% of our data to be twice annotated. Each of the annotators had to label 4 transcripts at this stage, one from each company. At the end, they meet and discuss disagreements to proceed the final version (gold annotation) on this part. The value of this

[3] https://site.financialmodelingprep.com/developer.
[4] https://labelstud.io/.
[5] https://github.com/Alaa-Ah/The-FinArgQuality-dataset-Quality-of-managers-arguments-in-Eearnings-Conference-Calls.

strategy is that it guarantees direct discussions between every pair of annotators, which help at the end to unify their mindset towards the annotations.

We report in the following Cohen's kappa inter-annotator agreement [39] on our *FinArgQuality* final corpus. For all data, we measure the agreement separately for each pair of annotators, and report the average. Table 2 shows that we obtained fair to substantial agreements [40].

Table 2. Inter-annotator agreement of the overall argument quality and unit types

Company	Specific	Persuasive	Strong	Objective	Temporal-history	Claim (All types)	Premise (All types)
MSFT	0.63	0.64	0.72	0.65	0.79	0.61	0.59
FB	0.33	0.13	0.21	0.36	0.66	0.56	0.57
AAPL	0.31	0.31	0.35	0.27	0.55	0.66	0.69
AMZN	0.11	0.21	0.24	0.36	0.26	0.37	0.51
All	0.345	0.322	0.38	0.41	0.565	0.55	0.59

We compare our results to a similar study by Wachsmuth et al. [41], that introduced the *Dagstuhl15512 ArgQuality Corpus* for ranking argumentation quality based on their developed taxonomy of 15 dimensions. They also adopted a 3-point scale (low, medium, high) for rating. Therefore, we consider this data as the most relevant to compare with. They reported Krippendorf's α of all annotators ranging from 0.174 to 0.447 only.

By analyzing disagreements, we found that the main source of disagreement is the missing of unit boundaries (Speech-To-Text nature of the transcripts' data), and the multiple possible interpretations of argument structure [42–44]. This, definitely, applies to rating argument quality, which is even more inherently subjective [11]. In addition, a high proportion of disagreement is associated with arguments that include modal verbs, and uncertainty quantification (e.g., "many", "some") which may hastily perceived with low degrees of specificity, strength, and persuasiveness. Thus, extending guidelines with those cases would improve further annotations.

4.3 *FinArgQuality* Data Statistics

The overall quality dimensions are described in Table 3 in total, and per company. The percentages are based on the total number of arguments. Overall, the score 1 is always the most associated with specific, persuasive and strong quality dimensions. Argument objectivity is validated mostly when mentioning unbiased indicators, such as numerical values or time references. We also notice that label 0 (low) is the least frequent. In addition, only 0.4% of the arguments are considered bad, i.e., all four dimensions (specific, persuasive, strong and objective) are rated by zero. This small percentage reflects the overall good quality

Table 3. Statistics of overall argument quality dimensions over *FinArgQuality*

Dimension	Company									
	FB		AMZN		MSFT		AAPL		Total	
	Count	[%]	Count	[%]	Count	[%]	Count	[%]	Count	[%]
Specific 0	29.0	1.33	13.0	0.60	34.0	1.56	7.0	0.32	83.0	3.80
Specific 1	281.0	12.87	202.0	9.25	466.0	21.34	147.0	6.73	1096.0	50.18
Specific 2	180.0	8.24	220.0	10.07	309.0	14.15	296.0	13.55	1005.0	46.02
Strong 0	39.0	1.79	31.0	1.42	49.0	2.24	19.0	0.87	138.0	6.32
Strong 1	317.0	14.51	274.0	12.55	557.0	25.50	285.0	13.05	1433.0	65.61
Strong 2	134.0	6.14	130.0	5.95	203.0	9.29	146.0	6.68	613.0	28.07
Persuasive 0	70.0	3.21	20.0	0.92	37.0	1.69	11.0	0.50	138.0	6.32
Persuasive 1	254.0	11.63	209.0	9.57	370.0	16.94	221.0	10.12	1054.0	48.26
Persuasive 2	166.0	7.60	206.0	9.43	402.0	18.41	218.0	9.98	992.0	45.42
Objective 0	102.0	4.67	76.0	3.48	304.0	13.92	149.0	6.82	631.0	28.89
Objective 1	388.0	17.77	359.0	16.44	505.0	23.12	301.0	13.78	1553.0	71.11
Temp.history -1	338.0	15.48	288.0	13.19	733.0	33.56	408.0	18.68	1767.0	80.91
Temp.history 0	26.0	1.19	18.0	0.82	12.0	0.55	4.0	0.18	60.0	2.75
Temp.history 1	54.0	2.47	43.0	1.97	7.0	0.32	10.0	0.46	114.0	5.22
Temp.history 2	24.0	1.10	41.0	1.88	17.0	0.78	11.0	0.50	93.0	4.26
Temp.history 3	48.0	2.20	45.0	2.06	40.0	1.83	17.0	0.78	150.0	6.87

of arguments, and the persuasion strategies which managers often use during the earnings calls and public speech, as highlighted by Crawford [33]. The time reference itself, is defined in our guidelines only in the past, as the temporal-history dimension. To standardize the annotations, we asked the annotators not to assume their interpretations of time references if it is not explicitly mentioned. Therefore, we got a majority class of -1, while all expressed time indicators compose about 20% of our arguments.

Furthermore, Table 4 exhibits the detailed corpus size and sentence/tokens distributions, as well as statistics with respect to the claim and premise types and argument relation. We can see that 43% of claims are factual, while 36% are based on opinions. The remaining claim types (Reformulated, Policy, Value, and Other) represent approximately 21%. This is reasonable since managers mainly report facts, or explain their views and future prospects.

Moreover, the distribution of premise types confirms the financial nature of the collected data, since it mostly covers facts (71%) and statistics (13%). Nevertheless, some background information seems to be annotated as facts by our annotators, given that it is still true (happened) information that could be tricky not to consider as a fact. Similarly, Carlile et al. [30] found that 493 of their premises received a "common_knoweledge" label, out of 707 premises with 8 potential premises types defined in the annotation guidelines. In a related analysis, Villalba and Saint-Dizier [35] show how "a number of evaluative expressions with a'heavy' semantic load receive an argumentative interpretation".

Table 4. Size and statistics of argument components types and argument relation over *FinArgQuality*. The average is presented along with its standard deviation

	Attribute	Count	[%]	Avg. per doc	Avg. per company
Sentences	In-argument	9693	68.53	12 ± 6	2423 ± 423
	Out-of-argument	4453	31.47	6 ± 4	1113 ± 158
Tokens	In-argument	244253	78.84	297 ± 155	61063 ±13437
	Out-of-argument	65537	21.16	82 ± 78	16384 ± 4796
Arg. components	Premises	5078	52.40	6 ± 4	1270 ± 271
	Claims	4613	47.60	6 ± 3	1153 ± 158
Claims	Fact	2001	43.38	3 ± 2	500 ± 93
	Opinion(view)	1672	36.25	2 ± 2	418 ± 64
	Reformulated	850	18.43	2 ± 1	212 ± 56
	Policy	45	0.99	1 ± 0	11 ± 5
	Value	28	0.60	1 ± 0	7 ± 3
	Other	17	0.37	1 ± 0	4 ± 3
Premises	Fact	3624	71.37	5 ± 3	906 ± 303
	Statistic	691	13.60	2 ± 1	173 ± 92
	Real Example	496	9.77	2 ± 1	124 ± 53
	Hypothesis	46	0.91	1 ± 0	12 ± 5
	Other	221	4.35	2 ± 1	55 ± 24
Relation types	Support	4823	98.41	6 ± 4	1206 ± 271
	Attack	78	1.59	1 ± 1	20 ± 10

5 Discussion and Conclusions

Recently, financial argumentation gained momentum in different languages (e.g., [45,46]). Given that both financial NLP and computational argumentation communities suffer from the lack of labeled data, we believe that our proposed assessment model and publicly available dataset, can serve as strong baselines for future work. We expect our carefully developed corpus to prompt various directions.

On the one hand, six potential argument mining tasks could be investigated using our data: argument identification, argument unit classification, argument relation classification, premise type multi-class classification, claim type multi-class classification, and overall argument quality assessment. In addition, a future direction could be to use the argumentative unit types in order to mine the argumentation strategies [47].

On the other hand, the automatic detection and qualification of arguments in financial domain is important for several goals, including but not limited to:

– Efficiency: It allows for the analysis of large amounts of financial data and reports quickly and accurately, reducing the time and resources required to manually identify and analyze arguments.

- Objectivity: It eliminates the subjective biases that can occur when humans are manually reviewing financial data, resulting in a more objective analysis.
- Enhanced market transparency: Arguments can provide more visibility into the reasoning behind investment decisions or analysts' recommendations, improving market transparency and trust.
- Improved decision-making: Providing a comprehensive analysis of financial arguments, identifying worthiness, and detection of verified claims, can help inform and improve decision-making in finance. For example, by using argument mining to validate managers' claims (e.g., [9]), we can assess the objectivity, completeness, and credibility of these arguments, providing investors with more informed and reliable insights for making investment decisions.

Last but not least, we would like to note that a first emerging output of this work is established by the FinArg-1 Shared Task[6], in cooperation with AIST, Japan[7], as well as different other partners. FinArg–1 covers argument unit and relation classification. The data reported in this paper is planned to be used in the next editions of this task. We plan to cover different tasks, including argument quality assessment. Our ultimate goal is to improve the automatic understanding of financial text. Therefore, we hope that the work presented in this paper fuels and inspires more research in computational argumentation, stock market and their interplay.

Acknowledgement. I would like to thank, for most, Prof. Ralf Kellner and Mr. Lukas Marx, for their open discussions and insightful feedback, which were essential to refine those argument quality dimensions from a financial point of view. In addition, I would like to thank all our annotators, including M. Kürsad Lacin.

References

1. Philippon, T.: The fintech opportunity. Technical report, National Bureau of Economic Research (2016)
2. Keith, K.A., Stent, A.: Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. arXiv preprint arXiv:1906.02868 (2019)
3. Matsumoto, D., Pronk, M., Roelofsen, E.: What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. Account. Rev. **86**(4), 1383–1414 (2011)
4. Price, S.M., Doran, J.S., Peterson, D.R., Bliss, B.A.: Earnings conference calls and stock returns: the incremental informativeness of textual tone. J. Bank. Financ. **36**(4), 992–1011 (2012)
5. Ma, Z., Bang, G., Wang, C., Liu, X.: Towards earnings call and stock price movement. arXiv preprint arXiv:2009.01317 (2020)
6. Li, J., Yang, L., Smyth, B., Dong, R.: MAEC: a multimodal aligned earnings conference call dataset for financial risk prediction. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3063–3070 (2020)

[6] http://finarg.nlpfin.com/.
[7] https://www.aist.go.jp/index_en.html.

7. Ye, Z., Qin, Y., Xu, W.: Financial risk prediction with multi-round Q&A attention network. In: IJCAI, pp. 4576–4582 (2020)
8. Palmieri, R.: The role of argumentation in financial communication and investor relations. In: Handbook of Financial Communication and Investor Relations, pp. 45–60 (2017)
9. Stenvall, J.: Management earnings forecasts: could an investor reliably detect an unduly positive bias on the basis of the strength of the argumentation? J. Bus. Commun. **48**(4), 393–408 (2011)
10. Alhamzeh, A., Fonck, R., Versmée, E., Egyed-Zsigmond, E., Kosch, H., Brunie, L.: It's time to reason: annotating argumentation structures in financial earnings calls: the FinArg dataset. In: Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), Abu Dhabi, United Arab Emirates (Hybrid), pp. 163–169. Association for Computational Linguistics (2022)
11. Wachsmuth, H., et al.: Argumentation quality assessment: theory vs. practice. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, pp. 250–255. Association for Computational Linguistics (2017)
12. Persing, I., Ng, V.: Modeling argument strength in student essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 543–552 (2015)
13. Habernal, I., Gurevych, I.: What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1214–1223 (2016)
14. Farra, N., Somasundaran, S., Burstein, J.: Scoring persuasive essays using opinions and their targets. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 64–74 (2015)
15. Rinott, R., Dankin, L., Perez, C.A., Khapra, M.M., Aharoni, E., Slonim, N.: 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, Lisbon, Portugal, pp. 440–450. Association for Computational Linguistics (2015)
16. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. In: Proceedings of the 25th International Conference on World Wide Web, pp. 613–624 (2016)
17. Stab, C., Gurevych, I.: Recognizing insufficiently supported arguments in argumentative essays. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 980–990 (2017)
18. Lauscher, A., Ng, L., Napoles, C., Tetreault, J.: Rhetoric, logic, and dialectic: advancing theory-based argument quality assessment in natural language processing. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), pp. 4563–4574. International Committee on Computational Linguistics (2020)
19. Wachsmuth, H., Werner, T.: Intrinsic quality assessment of arguments. arXiv preprint arXiv:2010.12473 (2020)
20. Gurcke, T., Alshomary, M., Wachsmuth, H.: Assessing the sufficiency of arguments through conclusion generation. arXiv preprint arXiv:2110.13495 (2021)

21. Walton, D.N.: Argument structure: a pragmatic theory. University of Toronto Press, Toronto (1996)
22. Chen, C.-C., Huang, H.-H., Chen, H.-H.: Evaluating the rationales of amateur investors. In: 2021 Proceedings of the Web Conference, pp. 3987–3998 (2021)
23. Zong, S., Ritter, A., Hovy, E.: Measuring forecasting skill from text. arXiv preprint arXiv:2006.07425 (2020)
24. Qin, Y., Yang, Y.: What you say and how you say it matters: predicting stock volatility using verbal and vocal cues. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 390–401 (2019)
25. Alhamzeh, A., et al.: A hybrid approach for stock market prediction using financial news and stocktwits. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 15–26. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_2
26. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: liwc and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
27. Rocci, A., Raimondo, C., Puccinelli, D.: Evidentiality and disagreement in earnings conference calls: preliminary empirical findings, pp. 100–104 (2019)
28. Prokofieva, A., Hirschberg, J.: Hedging and speaker commitment. In: 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland (2014)
29. Johnson, R.H., Blair, J.A.: Logical self-defense. In: International Debate Education Association (2006)
30. Carlile, W., Gurrapadi, N., Ke, Z., Ng, V.: Give me more feedback: annotating argument persuasiveness and related attributes in student essays. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 621–631. Association for Computational Linguistics (2018)
31. Chen, C.-C., Huang, H.-H., Shiue, T.-T., Chen, H.-H.: Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 136–143. IEEE (2018)
32. UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., Pustejovsky, J.: SemEval-2013 task 1: tempeval-3: evaluating time expressions, events, and temporal relations. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 1–9 (2013)
33. Camiciottoli, B.C.: Persuasion in earnings calls: a diachronic pragmalinguistic analysis. Int. J. Bus. Commun. **55**(3), 275–292 (2018)
34. Bentahar, J., Moulin, B., Bélanger, M.: A taxonomy of argumentation models used for knowledge representation. Artif. Intell. Rev. **33**(3), 211–259 (2010)
35. Villalba, M.P.G., Saint-Dizier, P.: Some facets of argument mining for opinion analysis. COMMA **245**, 23–34 (2012)
36. Khatib, K.A., Wachsmuth, H., Kiesel, J., Hagen, M., Stein, B.: A news editorial corpus for mining argumentation strategies. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3433–3443 (2016)
37. Al-Khatib, K.: Computational analysis of argumentation strategies. Dissertation, Bauhaus-Universität Weimar (2019)
38. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008)
39. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)

40. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 159–174 (1977)
41. Wachsmuth, H., et al.: Computational argumentation quality assessment in natural language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 176–187 (2017)
42. Walton, D., Reed, C.: Diagramming, argumentation schemes and critical questions. In: Van Eemeren, F.H., Blair, J.A., Willard, C.A., Snoeck Henkemans, A.F. (eds.) Anyone Who Has a View. Argumentation Library, vol. 8, pp. 195–211. Springer, Dordrecht (2003). https://doi.org/10.1007/978-94-007-1078-8_16
43. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1501–1510 (2014)
44. Henkemans, A.: State-of-the-art: the structure of argumentation. Argumentation **14**(4), 447–473 (2000)
45. Chen, C.-C., Huang, H.-H., Chen, H.-H.: From Opinion Mining to Financial Argument Mining. Springer, Heidelberg (2021). https://doi.org/10.1007/978-981-16-2881-8
46. Fishcheva, I., Osadchiy, D., Bochenina, K., Kotelnikov, E.: Argumentative text generation in economic domain. arXiv preprint arXiv:2206.09251 (2022)
47. Al-Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., Stein, B., Göring, S.: Webis-editorials-16 (2016)