# It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset

**Alaa Alhamzeh**[1,2*], **Romain Fonck**[1,2+], **Erwan Versmée**[1,2+], **Előd Egyed-Zsigmond**[1], **Harald Kosch**[2] and **Lionel Brunie**[1]
[1]INSA de Lyon, France
[2]Universität Passau, Germany
{firstname.secondname}@uni-passau.de, {firstname.secondname}@insa-lyon.fr

## Abstract

With the goal of reasoning on the financial textual data, we present in this paper, a novel approach for annotating arguments, their components and relations in the transcripts of earnings conference calls (ECCs). The proposed scheme is driven from the argumentation theory at the micro-structure level of discourse. We further conduct a manual annotation study with four annotators on 136 documents. By that, we obtained inter-annotator agreement of $\alpha_U$ = 0.70 for argument components and $\alpha$ = 0.81 for argument relations. The final created corpus, with the size of 804 documents, as well as the annotation guidelines are publicly available for researchers in the domains of computational argumentation, finance and FinNLP.

## 1 Introduction

The rise of data and the development of machine learning have led to the interdisciplinary financial technology (FinTech) that aims at supporting the financial industry with digital innovations and technology-enabled business models [Philippon, 2016]. Different applications have been explored such as fraud detection, digital payment, blockchain and trading systems. In terms of the latter, several factors impact its movement and it is hard, in reality, to get a very accurate prediction of the future stock prices. The *efficient market hypothesis* theory [Fama, 1970] states that it is impossible to "beat the market" consistently since current prices incorporate all available information and expectations. Nevertheless, the current view of the market comes from behavioural economics which see humans as irrational beings who are influenced by biases and experience when making investment decisions. In our previous work [Alhamzeh *et al.*, 2021b], we analysed the impact of stockTwits [1] and online news using a hybrid approach which consists of sentiment and event-based features as well as the price information for different observation and prediction time windows. [Chen *et al.*, 2021a] aimed at capturing expert-like rationales from social media platforms without the requirement of the annotated data. Similarly, [Zong *et al.*, 2020] hit the question: what makes some forecasters better than others? By exploring connections between the language people use to describe their predictions and their forecasting skills. On the other hand, [Keith and Stent, 2019] targeted the prediction of professional analysts recommendations who influence the decisions of many investors towards buying or selling in particular markets. Their findings confirm that earnings calls are moderately predictive of analysts' decisions even though theses decisions are function of different parameters including private communication with company executives and market conditions.

Moreover, while different works considered the sentiment and semantic analysis of text, we are looking towards a deeper understanding and interpretation of the language by the means of argument mining. According to [Chen *et al.*, 2021b], argument mining can be applied to understand the public's expectations for the market, providing valuable information for investment and other close applications. While they mostly studied the investors' posts on social platforms, we aim to particularly study the impact of arguments on the professional analysts themselves during the earnings conference calls (ECCs). Therefore, we present in this paper the first step of that methodology by discovering and annotating argumentation structures in ECCs.

ECCs are generally held in every fiscal quarter and consist of three main parts: a safe harbor statement, a presentation and the question answering (Q&A) session. In the presentation, executives give the statements about the performance of company in the last quarter and exhibit their expectations for the next one. During the Q&A session, professional analysts ask their questions and demand clarifications from the company's representatives. Different studies found that the discussion during the question answering session is the most informative and influencing part on the market [Matsumoto *et al.*, 2011; Price *et al.*, 2012]. Therefore, we focus in our study on these sessions, and more specially on annotating the arguments, their components and relations in the given answers of the company executives, where they try to justify their opinions and convince the other party to believe in them, which is indeed the essence of argumentation [Alhamzeh *et al.*, 2021a].

To the best of our knowledge, no prior work has been carried out to annotate arguments in earnings calls transcripts.

*Contact Author, + Equal Contribution
[1]https://stocktwits.com/



*Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163 - 169

Therefore, the contributions of this paper are the following:

- First, we introduce a novel annotation scheme for modeling arguments in the answers of Q&A sections of earnings calls conferences.
- Second, we present our annotation study and the reliability of data by the inter-annotator agreement with four annotators.
- Third, we evaluate our data on using a fine-tuned DistilBERT model [Sanh *et al.*, 2019] for the argument identification task.
- Fourth, we provide our annotated FinArg corpus freely to encourage future research [2].

This paper is organized as follows: in Section 2, we explore a conceptual background of argumentation modeling, and we highlight related works on argumentation in finance and on ECCs in particular. In Section 3, we present our proposed annotation scheme to model the argument components and relations in the executives' answers stated during the earning call. We further illustrate in Section 4 the whole process of corpus creation. We move to DistilBERT results on our dataset in Section 5. We finally move back to the big picture of the financial application and discuss the future directions in Section 6.

## 2 Related Work

### 2.1 Argumentation Models: Background

Argumentation is a fundamental aspect of human communication, thinking, and decision making [Alhamzeh *et al.*, 2021a]. The simplest form of argument consists of one premise (also known as evidence or reason) supporting one claim (also known as a conclusion). Therefore, recognizing arguments in text includes several subtasks [Stab and Gurevych, 2014]: (1) argument identification by separating argumentative from non-argumentative text units. (2) argument unit classification by further identifying premises and claims in the argumentative units, and (3) argument structure identification to associate relations between argument components.

Since the study of argumentation involves philosophy, logic, communication science, and more recently computer science, the literature reports a diverse range of proposals to model argumentation based on the text genre and the task at hand. [Bentahar *et al.*, 2010] organized arguments models into three categories:

- Monological models: focus on the internal structure of an argument (micro-structure).
- Dialogical models: focus on the relations between arguments in a discussion, debate or similar (macro-structure).
- Rhetorical models: focus on the rhetorical patterns of arguments (neither micro nor macro-structure).

[2] https://github.com/Alaa-Ah/The-FinArg-Dataset-Argument-Mining-in-Financial-Earnings-Calls

Those three perspectives on the study of argumentation are closely related [Walton and Reed, 2003]. In our study, we focus on the monological perspective, which is more relevant to our data type and well-suited for developing computational method [Peldszus and Stede, 2013]. Toulmin's model [Toulmin, 2003] is a well known argument model that formalize the internal micro-structure of an argument optimally by means of six parts: claim, data, warrant, backing, qualifier and rebuttal. [Chen *et al.*, 2021a] proposed to use this model to structure argumentation in analysts' opinions (in analysts' reports). However, this model has several drawbacks to model the daily life argumentation [Habernal and Gurevych, 2017a; Palau and Moens, 2009], mainly due to the fuzzy distinction between the defined argument components. For instance, the distinction between data, warrant and backing is often vague in practice [Freeman, 2011]. Therefore, we do not follow this model and instead we design a simpler annotation scheme which we will discuss in details in Section 3.

### 2.2 Earnings Conference Calls (ECCs)

The analysis of financial textual data has been studied from multiple aspects and types of documents in the state of the art. In terms of earning calls, one inspiring work is the study of [Keith and Stent, 2019] who identified a set of 20 pragmatic features of analysts' questions (e.g., hedging, concreteness and sentiment) during the earning conference calls which they correlate with analysts' pre-call investor recommendations. They also analyze the degree to which semantic and pragmatic features from an earnings call complement market data in predicting analysts' post-call changes in price targets. [Matsumoto *et al.*, 2011; Price *et al.*, 2012] found that the question-answer portions of earnings calls to be most informative. Moreover, given that executives cannot predict analysts' questions with complete certainty, executives' responses tend to be more unscripted than in the presentation section. Therefore, in our work, we focus only on Q&A sessions especially that it implies also the interaction with the analysts who we seek to understand their persuasion and decision making process via argumentation at the first place. In other words, we investigate only on the arguments stated in the answers of company representatives to the questions of professional analysts.

### 2.3 Argumentation in Finance

Argumentation in financial domain has been addressed mainly in communication studies in the literature [Palmieri, 2017; Hursti, 2011; Estrada and others, 2010]. Recently, [Pazienza *et al.*, 2019] introduced an abstract argumentation approach for the prediction of analysts' recommendations following earnings conference calls. They actually did not apply any argument mining method. Instead, they abstractly considered each question and answer as an argument, and they applied sentiment analysis between them to be considered as the relation itself.

On the other hand, there are huge efforts in the FinNLP domain, presented by Chen et al. [Chen *et al.*, 2021b]. However, most of their work efforts are towards the Chinese language (and market) while we consider mainly the English

language with respect to S&P 500. Furthermore, they have also organized a series of FinNum tasks that consider the numerical understanding with respect to the financial text properties. The challenge of 2021, namely, FinNum-3 [3] considers the classification of *in-claim* and *out-of-claim* numerals in the manager's speech during the earning call [Chen *et al.*, 2022]. However, this data answers only if a numeral is playing a role in a claim or not, without any extra information about premises or non-argumentative sentences. Moreover, sine one sentence may have two different labels of numerals (in and out of claim), we cannot know if this sentence represents a claim or not. In other words, the data is not about argument units, rather the focus is on the numeral understanding itself [Alhamzeh *et al.*, 2022].

Based on those studies and on our own experiments on different types of text, we found that ECCs are the best candidate for an argument-based solution. This could be justified by different reasons like the fact that social media posts are restricted with a maximum character count, and people tend to express their opinions and views more that structuring them in sort of premises and claims. For example, according to our analysis on stockTwits, different posts are only claims with no premises. Therefore, we build henceforth on the ECCs and we present the annotation study in the following sections.

## 3 Argumentation Structure in Earnings Calls

In this section, we discuss our proposed annotation scheme to model the argument components as well as the argumentative relations that constitute the argumentative discourse structure of earnings calls. We have first to point that the answers do not exhibit any common structure among all of them, to be hence structured as a connected tree or graph with circular relations. Rather, the answers are full of arguments that may or may not be directly related. This could be justified by the fact that those answers are part of an oral argumentation limited by time. Therefore, the company representatives tend to basically enumerate their evidences (premises) that support their claims. They may make the link between different claims and reasons they mentioned (or reformulate the same claim as well), whereas in most cases, they move to the next question. Hence and to simplify the task enough, we did not ask the annotators to define the relations between the arguments (macro-structure level) or to follow more fine-grained annotation scheme that will differentiate the major claim from other claims as in [Stab and Gurevych, 2014]. Instead, we are interested in detecting the arguments themselves as independent units. In particular, we model the structure of *each argument* using *one-claim-approach* proposed by [Cohen, 1987]. This approach considers only the root node of an argument as a claim and the remaining nodes in the structure as premises. The arrow from the premise to the claim composites the relation which can be either a support or an attack relation. Figure 1 represents a sample of our annotation scheme which implies that we can have different types of micro-structure arguments (e.g., basic, convergent and serial) in one answer.

[3] https://sites.google.com/nlg.csie.ntu.edu.tw/finnum3/task-definition?authuser=0

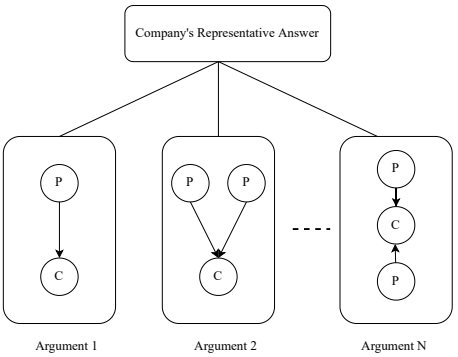


Figure 1: Argument annotation scheme (a sample) including argument components and argumentative relations (support/attack) indicated by arrows

Moreover, you can see a real example of the data in Figure 2. As we have mentioned, we are in particular interested in annotating the arguments stated by the company representatives. Therefore, in the answer of *Luca Maestri*, we see first some general information that is not argumentative (stated in Italic face), then the speaker start to argue about his claim ($C_1$) by stating different premises. The annotator marked every sentence ($P_1$ to $P_4$) as a premise since they all emphasize the stance of the speaker. In this example, all those premises belong to the same claim and they are all marked with a support relation type.

## 4 Corpus Creation

The motivation for creating a new corpus is threefold:

- First, we believe that it's time to reason the financial data and to move from shallow linguistic features and opinion mining to the reasons behind it, the analysis of persuading and decision making process via argument mining.
- Second, the lack of publicly available datasets is one of the big issues for the researchers who focus on both NLP and finance [Chen *et al.*, 2020].
- Third, the same challenge applies for argumentation field, where available datasets are often of small size and very domain and task dependent [Habernal and Gurevych, 2017b]. Therefore, our dataset can serve the computational argumentation scholars as well.

### 4.1 Data

We downloaded our data using the Financial Modeling Prep API [4]. We used Label Studio [5] as an annotation tool.

Our annotated data concerns the quarterly earnings calls of four companies: Amazon, Apple and Microsoft and Facebook for the period of 2015-2019 (i.e., we have 80 earning call transcripts). For each transcript, we created the list of all the speakers. After having determined the role of each of them (Analyst, Representative or an Operator), we were able

[4] https://site.financialmodelingprep.com/developer
[5] https://labelstud.io/

to split the whole text into different documents. Each document contains one or two questions asked by a single analyst and the corresponding response(s) by the company's representatives. We formulated these documents following Label Studio guidelines, and imported them to be further labeled with the argument units and relations.

In other words, we have a set of documents equal to the number of questions for each earning call, as far as every analyst asks only one question. In most cases, the same analyst will have two questions and two (or more) answers in the same document. Therefore, we observe a difference between the number of documents, number of questions, and the number of answers in our final corpus (cf. Table 1).

### 4.2 Argument Unit Segmentation

In the basic case, an argument component would be one complete sentence. However, in some cases, a sentence may contain several argument components. Accordingly, we annotated argument components at the clause-level (at minimum) and at the sentence level (at maximum). In other words, if we have complete statements in the same sentence we only consider them as different argument components if there is an inference relation between them. Particularly, neither statements connected with conjunctions like "and" or "or " nor conditional sentences (if, then) imply an inference relation. On the contrary, inference could appear in the following forms:

*"**claim** because of **premise**"*
*"Since **premise** then **claim**."*
*"In view of the fact **premise** that it follows that **claim**"*

However, since there is no punctuation in spoken language, segmentation is more challenging and it must be based on breaks, pitch, etc. In our case, we let the annotators segment each sentence based on the context with respect to the splitting roles we defined earlier.

### 4.3 Annotation Study

Our annotation study consists of three stages:

1. Annotation guidelines: We conduct a preliminary study to define the annotation guidelines with one of our annotators.
2. Pilot annotations: The goal of this stage was to test the annotation guidelines before a complete corpus is annotated. This was done by training sessions and discussions with the annotators. We got feedback from them to update the guidelines and solve unclear situations. We observed at this step that the annotation is more complicated in practice and even with our simple annotation scheme, one quarter takes between 2 to 3 hours to be completely annotated. This confirms our choice of annotation at the micro-structure level of argument and with the one-claim-approach.
3. Inter annotator agreement: To compute how homogeneous and thus reliable the annotations are.

### 4.4 Inter-Annotator Agreement (IAA)

To evaluate the reliability of our data, we determine a group of 12 earnings calls that represent about 20% of the whole data and covering all four companies to be annotated by a permutation of two annotators (out of four) separately. Those individual versions of the annotations are used later to compute the inter annotator agreement. To this end, we used Krippendorff's $\alpha_U$ [Krippendorff, 2004] and Krippendorff's $\alpha$ [Krippendorff, 1980] for the argument components and argument relations annotations respectively. That's because the former considers the differences in the markable boundaries of the two annotators and thus allows for assessing the reliability of argument units annotations. However, in terms of the relation annotation, the markables are the set of premise-claim pairs. We obtained a degree of $\alpha_U$ =0.70 for argument components and $\alpha$=0.81 for argument relations. Hence, we conclude that the annotation of arguments in earnings calls is reliably possible. However, it can be tricky to get identical annotations given that the argument component types are strongly related (i.e., the annotation of a claim depends on its connected premises). Therefore, every permutation of two annotators had to meet and discuss their disagreement cases to produce the last validated document *(gold annotations)*. As a result, we discovered that the primary source of uncertainty is due to the missing of sentence boundaries and the connected context that covers multiple sentences. Therefore, we asked the annotators to read the entire question to identify the controversial topic before starting with the actual annotation task on the answer paragraph. Despite the fact that this approach is more time-consuming than a direct identification of argument components and relations, it yields to a more reliable annotated data. Furthermore, the understanding of the question will help to assess the quality of the arguments which we will address in our future work.

### 4.5 Creation of the Final Corpus: the FinArg Dataset

Once we extracted our data annotated using LabelStudio, the output file is a very long document in JSON format. However, before using this data, we ran some scripts to detect annotations that did not follow the guidelines. In most of the cases, a document was classified wrong because it contained at least one of these three issues: the answer part of the document was not fully annotated, the same piece of text was annotated twice or a relation was misdirected. When it is possible, the issue was corrected by code. Otherwise, we ask the corresponding document's annotator to correct that mistake. Thereafter and to increase the usability and reproducibility of our FinArg dataset, we structured the important information of arguments in a similar way to the student essays dataset [Stab and Gurevych, 2014] since it is simply understandable and almost the most used one in computational argumentation.

Hence, the annotation document includes for every premise, claim or Non-argument text:

***"Id, label, start index, end index, text"***

and for every argument component relation:

***"Id, label, ARG1: source component id, ARG2: target component id"***

Moreover, we provide an additional JSON file including the following labels:

Operator (Intro): From JPMorgan, Rod Hall.
Rod Hall (Question): Hi, guys. Thanks for taking my questions. I wanted to start off just going back to the 165 million subscriptions and ask Tim or Luca if you could comment on the unique number of users there. And I think you had made a comment, Tim, in your prepared remarks that the average revenue per user was up, or maybe that was you, Luca. But if you guys could just talk about any more color around that average revenue per user, it would be interesting to us. And then I have one follow-up to that. Thanks.
Luca Maestri (Answer): *Yes, I'll take it, Rod. We don't disclose into the number of subscriptions. Of course, we're just giving you the total count of subscriptions that are out there. Of course, there are several customers that subscribe to more than one of our services.*
[ There is some level of overlap, but the total number of subscribers is very, very large, obviously less than 165 million ]$\mathbf{P}_1$.
[ But it's very good for us to see the breadth of subscriptions that we offer and that customers are interested in ]$\mathbf{C}_1$. *It is very large.*[ And if you remember, we quoted the same number a quarter ago and we talked about 150 million ]$\mathbf{P}_2$
[ So when you think about a sequential increase of 15 million subscriptions from the December quarter to the March quarter, it really gives you a sense for the momentum that we have on our content stores ]$\mathbf{P}_3$.
[ It's quite impressive to add 15 million subscriptions in 90 days. ]$\mathbf{P}_4$. [.....]

Figure 2: An example of the Apple Q2 2017- the annotation covers the answer where the Italic text is for *Non-argument*, Claim is marked as $\mathbf{C}_1$ and Premises are marked with $\mathbf{P}_{count}$

***Operator, Analyst, Representative, Intro, Question, Answer***

This latter annotations could be useful especially for a financial application scenario.

### 4.6 Corpus Statistics

Table 1 shows statistics about our annotated data distributions. The number of documents represents the number of different analysts. However, usually an analyst asks two different questions. Also, for some questions, two of the company representatives will answer separately. Therefore, the number of annotated answers can be (and it is) more than number of questions and about twice the number of documents. The found proportion between claims and premises is also common in argumentation and confirms the findings of [Mochales and Moens, 2011; Stab and Gurevych, 2014] that claims are usually supported by several premises for ensuring a complete and stable standpoint. Additionally, the proportion between support and attack relations is normal, since discussing the opposite point of view (as a strategy to prevent any future criticism) is less commonly used in argumentation comparing to the direct supporting premises. There is also a couple of unlinked premises or claims in the data, mostly for "reformulated" claims since we ask our annotators not to link them again to the same premises as the original stated claim. In other words, we want to avoid counting them as new arguments. Furthermore, Table 2 shows a detailed version of the classes distributions per different companies.

| Type | Count | % |
|---|---|---|
| **Documents** | 804 | - |
| **Questions** | 1553 | - |
| **Answers** | 1777 | - |
| **Premises** | 4894 | 35.856% |
| **Claims** | 4478 | 32.808% |
| **Non-argument** | 4277 | 31.336% |
| **Support** | 4604 | 98.355% |
| **Attack** | 77 | 1.645% |
| **Unlinked** | 1778 | 18.971% |

Table 1: Corpus statistics and class distribution

## 5 Evaluation

As a base-line model, we fine-tuned DistilBERT [Sanh *et al.*, 2019] with our dataset on the argument identification task (i.e., argument/ non-argument classification) at the sentence-level. Table 3 shows that we got an accuracy of 0.84 and F1-score of 0.80, which are comparable to DistilBERT outcomes on the well known argumentation corpora: *Student essays* [Stab and Gurevych, 2014] and *User-generated web discourse* [Habernal and Gurevych, 2017a] presented in [Alhamzeh *et al.*, 2021a] and the BERT-based results presented in [Wambsganss *et al.*, 2020].

Hence, our primary findings suggest that we can automatically export further earnings conference calls annotations with a good degree of reliability using a supervised machine learning algorithm trained on our corpus. Based on that, we can reach the granularity of data needed for future work on the prediction of analysts' post-call recommendations.

## 6 Conclusions and Future Work

Recently, different cutting-edge technologies have been addressed in FinTech domain, including numeracy understanding, opinion mining and financial document processing. In this paper, we contribute to the (1) theory, (2) data and (3) evaluation aspect of argumentation structure in the financial domain by (1) proposing a micro-structure argumentation scheme for modeling arguments presented in analysts' responses during the earnings conference calls, (2) work-

| Type | FB | AAPL | AMZN | MSFT |
|---|---|---|---|---|
| **Documents** | 264 | 140 | 213 | 187 |
| **Questions** | 421 | 431 | 330 | 371 |
| **Answers** | 489 | 431 | 330 | 527 |
| **Premises** | 1722 | 1035 | 1010 | 1127 |
| **Claims** | 1423 | 1103 | 969 | 983 |
| **Non-argument** | 1332 | 1183 | 924 | 838 |
| **Support** | 1638 | 949 | 924 | 1093 |
| **Attack** | 20 | 35 | 6 | 16 |
| **Unlinked** | 385 | 499 | 457 | 437 |

Table 2: Distribution per company where FB: Facebook, AAPL: Apple, AMZN: Amazon, MSFT: Microsoft

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| DistilBERT | 0.84 | 0.83 | 0.78 | 0.80 |

Table 3: Evaluations of the DistilBERT fine-tuned model on the **FinArg** dataset

ing on the related annotation covering a period of five years (2015-2019) on four companies (FB, AMZN, MSFT, AAPL) to produce the FinArg dataset with a size 804 documents, and (3) evaluating this data using DistilBERT as a baseline model.

We aim in the future work to employ this data and train models to the end of prediction of analysts' post-call recommendations. This opens up different research questions like the required granularity of the data, the emission time of the recommendation's announcement, the analyst's questions (topic and sentiment) during the earning call (if applicable) and others. However, we believe that it's time to reason on financial textual data and to move from basic linguistic features, semantics and sentiment analysis to the reasons behind it and the quality of it with the help of argument mining and argument quality assessment which we will address in our future work as well. As a conclusion, we claim that our dataset presented in this paper will foster the research in FinTech domain in parallel with computational argumentation as an NLP task itself.

## References


[Alhamzeh *et al.*, 2021a] Alaa Alhamzeh, Bouhaouel Mohamed, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. A stacking approach for cross-domain argument identification. In *International Conference on Database and Expert Systems Applications*, pages 361–373. Springer, 2021.

[Alhamzeh *et al.*, 2021b] Alaa Alhamzeh, Saptarshi Mukhopadhaya, Salim Hafid, Alexandre Bremard, Előd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. A hybrid approach for stock market prediction using financial news and stocktwits. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 15–26. Springer, 2021.

[Alhamzeh *et al.*, 2022] Alaa Alhamzeh, M. Kürsad Lacin, and Előd Egyed-Zsigmond. Passau21 at the ntcir-16 finnum-3 task: Prediction of numerical claims in the earnings calls with transfer learning. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, pp. 121-125, 2022. Tokyo, Japan*, 2022.

[Bentahar *et al.*, 2010] Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259, 2010.

[Chen *et al.*, 2020] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Nlp in fintech applications: past, present and future. *arXiv preprint arXiv:2005.01320*, 2020.

[Chen *et al.*, 2021a] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, pages 3987–3998, 2021.

[Chen *et al.*, 2021b] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From Opinion Mining to Financial Argument Mining*. Springer Nature, 2021.

[Chen *et al.*, 2022] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-16 finnum-3 task: Investor's and manager's fine-grained claim detection. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*, 2022.

[Cohen, 1987] Robin Cohen. Analyzing the structure of argumentative discourse. *Computational linguistics*, 13:11–24, 1987.

[Estrada and others, 2010] Fernando Estrada et al. Theory of argumentation in financial markets. *Journal of Advanced Studies in Finance (JASF)*, 1(01):18–22, 2010.

[Fama, 1970] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[Freeman, 2011] James B. Freeman. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter Mouton, 2011.

[Habernal and Gurevych, 2017a] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.

[Habernal and Gurevych, 2017b] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017.

[Hursti, 2011] Kristian Hursti. Management earnings forecasts: Could an investor reliably detect an unduly positive bias on the basis of the strength of the argumentation? *The Journal of Business Communication (1973)*, 48(4):393–408, 2011.

[Keith and Stent, 2019] Katherine A Keith and Amanda Stent. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. *arXiv preprint arXiv:1906.02868*, 2019.

[Krippendorff, 1980] Klaus Krippendorff. Content analysis: An introduction to its methodology. *Sage*, 1980.

[Krippendorff, 2004] Klaus Krippendorff. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800, 2004.

[Matsumoto *et al.*, 2011] Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, 86(4):1383–1414, 2011.

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

[Palau and Moens, 2009] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, 2009.

[Palmieri, 2017] Rudi Palmieri. The role of argumentation in financial communication and investor relations. *Handbook of financial communication and investor relations*, pages 45–60, 2017.

[Pazienza *et al.*, 2019] Andrea Pazienza, Davide Grossi, Floriana Grasso, Rudi Palmieri, Michele Zito, and Stefano Ferilli. An abstract argumentation approach for the prediction of analysts' recommendations following earnings conference calls. *Intelligenza Artificiale*, 13(2):173–188, 2019.

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

[Philippon, 2016] Thomas Philippon. The fintech opportunity. Technical report, National Bureau of Economic Research, 2016.

[Price *et al.*, 2012] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Stab and Gurevych, 2014] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510, 2014.

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

[Walton and Reed, 2003] Douglas Walton and Chris Reed. Diagramming, argumentation schemes and critical questions. In *Anyone Who Has a View*, pages 195–211. Springer, 2003.

[Wambsganss *et al.*, 2020] Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. Unlocking transfer learning in argumentation mining: A domain-independent modelling approach. In *15th International Conference on Wirtschaftsinformatik*, 2020.

[Zong *et al.*, 2020] Shi Zong, Alan Ritter, and Eduard Hovy. Measuring forecasting skill from text. *arXiv preprint arXiv:2006.07425*, 2020.