**Doctoral Thesis**

# Language Reasoning by means of Argument Mining and Argument Quality

**Alaa Alhamzeh**

Defended publicly the 15th of May 2023, in front of the jury, composed of:

| | | |
|---|---|---|
| Jacques SAVOY | Professor, Université de Neuchâtel | Reviewer |
| Elisabeth LEX | Associate Professor HdR, Graz University of Technology | Reviewer |
| Chantal SOULE-DUPUY | Professor, Université de Toulouse | Examiner |
| Catherine BERRUT | Professor, Université de Grenoble | Examiner |
| Michael GRANITZER | Professor, Universität Passau | Examiner |
| Harald KOSCH | Professor, Universität Passau | Co-director |
| Elöd EGYED-ZSIGMOND | Associate Professor HdR, INSA-Lyon | Co-director |
| Lionel BRUNIE | Professor, INSA-Lyon, | Co-director |


The Cotutelle-PhD has been conducted within the framework of the *International Research & Innovation Center on Intelligent Digital Systems (IRIXYS)*.


The Cotutelle-PhD has been financially supported by the *Deutsch-Französische Hochschule (DFH) / Universitè franco-allemande (UFA)*.

Alaa Alhamzeh. *Language Reasoning by means of Argument Mining and Argument Quality*, May 2023

## ABSTRACT - ENGLISH

Understanding of financial data has always been a point of interest for market participants to make better informed decisions. Recently, different cutting-edge technologies have been addressed in the Financial Technology (FinTech) domain, including numeracy understanding, opinion mining and financial document processing.

In this thesis, we are interested in analyzing the arguments of financial experts with the goal of supporting investment decisions.

Although various business studies confirm the crucial role of argumentation in financial communications, no work has addressed this problem as a computational argumentation task. In other words, the automatic analysis of arguments. In this regard, this thesis presents contributions in the three essential axes of theory, data, and evaluation to fill the gap between argument mining and financial text.

First, we propose a method for determining the structure of the arguments stated by company representatives during the public announcement of their quarterly results and future estimations through earnings conference calls. The proposed scheme is derived from argumentation theory at the micro-structure level of discourse. We further conducted the corresponding annotation study and published the first financial dataset annotated with arguments: *FinArg*.

Moreover, we investigate the question of evaluating the quality of arguments in this financial genre of text. To tackle this challenge, we suggest using two levels of quality metrics, considering both the Natural Language Processing (NLP) literature of argument quality assessment and the financial era peculiarities.

Hence, we have also enriched the *FinArg* data with our quality dimensions to produce the *FinArgQuality* dataset.

In terms of evaluation, we validate the principle of ensemble learning on the argument identification and argument unit classification tasks. We show that combining a traditional machine learning model along with a deep learning one, via an integration model (stacking), improves the overall performance, especially in small dataset settings.

In addition, despite the fact that argument mining is mainly a domain-dependent task, to this date, the number of studies that tackle the generalization of argument mining models is still relatively small. Therefore, using our stacking approach and in comparison to the transfer learning model of DistilBert, we address and analyze three real-world scenarios concerning the model robustness over completely unseen domains and unseen topics.

Furthermore, with the aim of the automatic assessment of argument strength, we have investigated and compared different (refined) versions of Bert-based models that incorporate external knowledge in the decision layer. Conse-

quently, our method outperforms the baseline model by 13 ± 2% in terms of F1-score through integrating Bert with encoded categorical features.

Beyond our theoretical and methodological proposals, our model of argument quality assessment, annotated corpora, and evaluation approaches are publicly available, and can serve as strong baselines for future work in both FinNLP and computational argumentation domains.

Hence, directly exploiting this thesis, we proposed to the community, a new task/challenge related to the analysis of financial arguments: *FinArg-1*, within the framework of the NTCIR-17 conference.

We also used our proposals to react to the *Touché* challenge at the CLEF 2021 conference. Our contribution was selected among the « Best of Labs ».

# RÉSUMÉ

La compréhension des données financières a toujours été un point d'intérêt pour les participants au marché afin de prendre des décisions plus informées et pertinentes. Récemment, différentes technologies de pointe ont été abordées dans le domaine de la technologie financière (FinTech), comprenant l'analyse de numératie, l'analyse d'opinion et le traitement de documents financiers.

Dans cette thèse, nous nous intéressons à l'analyse des arguments défendus par les experts financiers en vue d'étayer une décision d'investissement.

Bien que diverses études de cas confirment le rôle crucial de l'argumentation dans les communications financières, aucun travail n'a abordé ce problème en tant que problématique computationnelle, c'est-à-dire d'analyse automatique des arguments. Centrée sur cette problématique, cette thèse apporte des contributions dans ses trois principales axes: théorie, mise en œuvre, et évaluation.

En premier lieu, nous proposons une méthode pour annoter la structure des arguments énoncés par les représentants d'entreprises lors de la publication de résultats financiers (« earnings conference calls »). Le schéma proposé est dérivé de la théorie de l'argumentation au niveau de la micro-structure du discours.

Nous avons également, dans ce cadre, mené une étude sur cette annotation et publié le premier jeu de données financières annoté avec des arguments: *FinArg*.

De plus, nous étudions en profondeur la question de l'évaluation de la qualité des arguments dans ce type de textes et proposons deux niveaux de métriques de qualité, en prenant en compte à la fois la littérature NLP sur la qualité des arguments et les particularités de la finance numérique.

Nous avons également enrichi les données *FinArg* avec ces dimensions de qualité pour produire le jeu de données *FinArgQuality*.

En termes d'évaluation, nous validons le principe de l'apprentissage ensembliste pour les tâches d'identification d'arguments et de classification d'unités d'argument. Nous montrons que la combinaison d'un modèle d'apprentissage automatique traditionnel avec un modèle d'apprentissage profond, via un modèle d'intégration (« empilement »), améliore les performances globales, en particulier dans le cas de petits ensembles de données.

Bien que l'extraction d'arguments soit principalement une tâche dépendante du domaine, à ce jour, le nombre d'études traitant de la généralisation des modèles d'extraction d'arguments est toujours relativement faible. En utilisant notre approche d'empilement et en comparant avec le modèle d'apprentissage par transfert de DistilBert, nous décrivons et analysons trois scénarios réels ciblant la robustesse du modèle sur des domaines et des thèmes non vus.

De plus, dans le but d'évaluer automatiquement la force d'argumentation, nous avons étudié et comparé différentes versions (affinées) de modèles basés

sur Bert. Les résultats obtenus dépassent ceux du modèle de base de 13 ± 2% en termes de score F1 en intégrant Bert avec des caractéristiques catégorielles codées.

Au-delà de nos propositions théoriques et méthodologiques, les dimensions d'évaluation de la qualité que nous avons proposées, les corpus annotés et les méthodes d'évaluation sont disponibles publiquement et peuvent servir de bases solides pour des travaux futurs dans les domaines de "FinNLP" et de l'analyse computationnelle d'argumentations.

Ainsi, exploitation directe de cette thèse, nous avons décrit et proposé à la communauté, dans le cadre de la conférence NTCIR-17, une nouvelle tâche/challenge, la tâche *FinArg-1*, portant sur l'analyse d'argumentations financières.

Nous avons également exploité nos propositions pour répondre au challenge *Touché* de la conférence CLEF 2021. Notre contribution a été sélectionnée parmi les « Best of Labs ».

CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS

**NLP** Natural Language Processing.

**AM** Argument Mining.
**AQ** Argument Quality.

**CAQ** Computational Argument Quality.

**FinTech** Financial Technology.
**FinNLP** Financial Natural Language Processing.

**IAA** Inter-Annotator Agreement.

**ECCs** Earnings Conference Calls.

**LLM** Large Language Models.

**BERT** Bidirectional Encoder Representations from Transformers.

**DistilBERT** a distilled version of BERT.

**RoBERTa** A Robustly Optimized BERT Pretraining Approach.

**XLNet** Generalized Autoregressive Pretraining for Language Understanding.

**GPT-3** Generative Pre-trained Transformer 3.

**PoS** Part of Speech.

**BoW** Bag of Words.

**NER** Named Entity Recognition.

**SVM** Support Vector Machines.

**QE** Query Expansion.

**SE** Student Essays dataset.

**WD** User Generated Web-Discourse dataset.

# 1

# INTRODUCTION

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 [Phi16]. Different applications have been explored such as fraud detection, digital payment, blockchain, and trading systems. However, given that about 80% of today's data is unstructured information which is composed mainly of textual data, it is important to leverage this data and to process it in different frameworks, which is called Financial Natural Language Processing (FinNLP).

Nevertheless, it is important to point out that FinNLP does not aim at replacing traditional methods like the fundamental analysis of stock market. Instead, it is a complementary approach where fundamental methods can be improved and augmented by the advent of NLP.

While previous studies have mainly investigated standard linguistic features and sentiment analysis, we aim, in this work, to mine and study the argumentative segments of a financial text. In particular, we investigate on the transcripts of financial Earnings Conference Calls (ECCs). 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 Q&A session where company representatives (mainly chief executive officer and chief financial officer) answer the questions of professional analysts and other market participants. However, many studies show that the question-answering session to be the most informative and impacting part on the market [MPR11; Pri+12; ma+20]. 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 understanding of earnings calls is valuable for different financial services and applications (e.g., financial risk prediction [Li+20; YQX20], modeling analysts' decision-making [KS19]). However, these calls are still an under-resourced text genre in computational argumentation. One may ask why do we need argument mining for analyzing earnings calls? Foremost, various business communication studies proved the important role of argumentation in ECCs (e.g., [Pal17; Hur11]). Yet, to date, the automatic detection and investigation of them is not feasible. Second, arguments can serve as a new reasonable tool for various financial applications and goals. For example, but not limited to:

- Argument-based opinion mining [CHC21b].
- Improve decision-making support systems, by the summarization of the long transcripts of ECCs, and highlighting the argumentative parts.
- Understanding correlations between the executives' quality of arguments and analysts' recommendations, or rather predicting the latter, as in [KS19].
- And similar.

In their book "From Opinion Mining to Financial Argument Mining", Chen et al. [CHC21b], claim that argument mining can be applied to understand the public's expectations of the market. They, therefore, studied mainly the investors' posts on social platforms. Moreover, they basically focus on the *Chinese* market (and Language). In addition, there are recent *Russian* attempts to investigate the financial argumentation through the study of Fishcheva et al. on financial argument generation [Fis+22].

Surprisingly, although the role of argumentation has been widely adopted in business communication and financial studies [Pal17], there are no empirical studies on its applicability to further financial goals. This interdisciplinary is relatively new. Therefore, we believe that a considerable part of our efforts was towards the *foundation* of this project, where many sessions of discussions were held with the Chair of Financial Data Analytics at the University of Passau[1].

Before we outline our research questions, we discuss the state of the art of financial argumentation in *English* and its drawbacks. To the best of our knowledge, there is only one attempt by Pazienza et al. [Paz+19] on English data. However, they assumed that each paragraph in the transcript of an earnings call to be a single "abstract" argument. This, nonetheless, does not correspond to the NLP definition of argument mining task. That is, the automatic detection and identification of argument structure (argument components and relations), where the simplest form of an argument is one *claim* supported by one *premise*.

Argumentation theory, which is mainly based on philosophical and linguistics studies, reports a rich fortune of argument models and more than 90 argument schemes [WRM08]. While the argument model studies the structure

1 https://www.wiwi.uni-passau.de/en/financial-data-analytics/research

of an argument, the scheme focuses on the inferential configuration that an argument uses. For instance, argument from analogy, argument from expert opinion, argument from example, and practical reasoning. Apparently, one argument may include more than one reasoning type and hence follow more than one argumentation scheme [WR03]. However, argumentation models mainly describe the relation between argument components, or with external arguments. The definition of what argument model to follow, depends on the data and the settings at first place [SG14b; HG17]. Hence, we had to study and analyze different proposals to decide which model is most relevant to our case of earnings calls.

Furthermore, another limitation of the work of Pazienza [Paz+19], is that no precise argument quality assessment could be done on the mere assignment of arguments as paragraphs. The quality of arguments (and for opinions in general) is crucial to estimate its usefulness to the audience (e.g., readers) [CHC21b].

Yet again, the definition of what is a good argument is very domain, and task-dependent [WW20; JB06]. Not surprisingly, Eemeren et al. [EH02] linked the speaker's strategy of maneuvering to the 'audience demand'. The market analysis has its peculiarities, which can be discounted using a normative argumentation.

Therefore, it is important to find the right trade-off, and to give both practical observations and theory standardization their due. For instance, while rhetorical figures may play a more considerable role in legal-text argumentation [Sau94], real-world values speak for themselves in the financial era. Therefore, we need to define the argument quality dimensions with respect to the features of this genre of text as well as the market perspective. We aim, in this thesis, at filling this research gap between computational argumentation and financial documents.

Our second contribution axis corresponds to the argument mining model evaluation. In fact, argumentation is a complex and nuanced topic that varies across different domains and disciplines, which makes argument mining algorithms and models often designed to be domain-specific. While having a highly specialized model is appropriate for some applications, developing a more flexible and adaptable model can be rather beneficial in real-world scenarios where the data may vary over time or between different users. Yet, developing a machine learning model that can mine the argumentative text or argument basic components (premises and claims) in cross-domain settings is still not fairly explored [MS16]. According to [Wan+22], developing more fair and application-driven evaluation standards, as well as the interpretability of the results, are one of the most challenging open issues in domain generalization.

We address the aforementioned challenges (i.e., argument mining model generalization, argument structure and argument quality assessment in ECCs) in more details along with the raised research questions in the following section.

## 1.1 RESEARCH QUESTIONS AND CONTRIBUTIONS

People argue differently, in their daily-life, social media, public debates, or official documents. The early work of the argument mining field targeted specific sources of data. For example, Stab et al. published the student essays corpus that labels argument structure in 402 essays [SG14a]. As a result, the corresponding parser is tailored to the particular language and discourse conventions of this specific domain. However, having highly homogenous data in a small size training set, raises the question whether the model is memorizing these examples rather than identifying patterns that are representative of the broader problem. According to [Zha+20a], domain generalization remains a major challenge for NLP systems.

On the one hand, training a generalized model can be more challenging, as it requires the model to learn and handle a larger and more diverse set of data [Wan+22]. On the other hand, the number of studies that address the evaluation of of the argument mining model's robustness to shifted domains, rather than to the test split, is still relatively small. This puts forward our first research question:

*RQ1* How to build and evaluate an argument parser that can generalize over heterogeneous corpora, given that argument mining is a domain-dependent task?

To tackle this question, we have inquired the robustness of the model generalization in terms of various argument mining tasks: argument identification and argument unit classification.

To get a better understanding of a model's generalization ability, we examine an ensemble learning approach that combines a traditional support vector machine (SVM) [CV95] with DistilBERT [San+19] model, in comparison to the fine-tuned transfer model of DistilBERT, through different testing scenarios on completely unseen data, unseen topics, and over different model runs.

*RQ2* How to model argumentation structure in earnings conference calls, given the diverse range of argument models and schemes proposed in argumentation theory, and can reliable annotations of the selected structure be achieved?

To answer this research question, we first provide a systematic review of the argumentation analysis strategy, and we study the proposed models in computational argumentation. Based on that, we propose an annotation scheme for annotating managers' arguments in the Q&A session of ECCs.

Given the complexity of the annotation task, and to provide good data quality, instead of the crowd annotations, the task was assigned to four annotators from computer science and economics disciplines. Consequently, we introduce **FinArg** dataset, the *first* financial argument mining corpus in English earnings calls.

We further examine the reliability of this data in terms of inter-annotator agreement, and our proposed ensemble learning model in *RQ1* on this data.

*RQ3* 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?

To answer this research question, we first review the existing proposals of assessing arguments quality as a computational argumentation task. Concurrently, we revise the literature of handling financial text. We make the link between those two lines of research by suggesting different argument quality metrics at both levels of "argument as one unit" and "argument components".

After creating the **FinArgQuality** dataset, we further reveal the found correlations between those metrics. Moreover, we build a deep learning model that automatically classify the strength score of a given argument.

In summary, this thesis aims at understanding argumentation approaches in finance, and moving the theories of business studies to a pragmatic and tangible use, by proposing annotation schemes (for argument structure and argument quality) and providing accessible data [2]. This data can be the fuel for various applications. In this thesis, we attempt to discover existing correlations between managers' quality of arguments and professional analysts' recommendations.

We have also provided a comprehensive study on machine learning models that are able to mine arguments from heterogeneous corpora. Finally, we suggest BERT with encoded features to the end of classifying strong arguments in ECCs.

Last but not least, during the work of this thesis, we have participated in two related activities: (1) Touché shared task 2021: Argument Retrieval for Comparative Question Answering[3]. (2) FinNum-3: Investor's and Manager's Fine-grained Claim Detection [4].

## 1.2 PUBLICATION RECORD

Most of the work presented in this thesis has been previously published in proceedings of peer-reviewed international journal, conferences, and workshops. This includes text, figures, and tables. We list them here in inverse chronological order.

• **Alaa Alhamzeh**, Előd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrović, Lionel Brunie and Harald Kosch. "Empirical Study of the Model Generalization for Argument Mining in Cross-domain and Cross-topic Settings". In: Transactions on Large-Scale Data-and Knowledge-Centered Systems (TLDKS Journal – Regular Papers), vol 13470. Springer, Berlin, Heidelberg, 2022, pp. 103–126.

• **Alaa Alhamzeh**, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie and Harald Kosch. "Query Expansion, Argument Mining

2 Most existing studies are based on private datasets e.g., Bloomberg [KS19].

3 https://touche.webis.de/clef21/touche21-web/argument-retrieval-for-comparative-questions

4 https://sites.google.com/nlg.csie.ntu.edu.tw/finnum3/finnum-3

and Document Scoring for an Efficient Question Answering System". In Best of 2021 Labs, International Conference of the Cross-Language Evaluation Forum for European Languages. Springer 2022 pp. 162-174., Bologna - Italy.
• **Alaa Alhamzeh**, Romain Fonck, Erwan Versmée, Előd Egyed-Zsigmond, Harald Kosch, Lionel Brunie. 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): Association for Computational Linguistics, Dec. 2022, pp. 163–169.
• **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. 2022.
• **Alaa Alhamzeh**, Mohamed Bouhaouel, 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. Springer. 2021, pp. 361–373.
• **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. Springer. 2021, pp. 15–26.
• **Alaa Alhamzeh**, Mohamed Bouhaouel, Előd Egyed-Zsigmond, and Jelena Mitrović. "DistilBERT-based Argumentation Retrieval for Answering Comparative Questions." In: Working Notes of CLEF (2021).
• The work of chapters 5 is under submission.
In addition to the publications listed before, which are directly related to the content of this thesis, the author contributed to and published other research work during the course of her doctoral studies:
• Régis Goubin, Dorian Lefeuvre, **Alaa Alhamzeh**, Jelena Mitrović, Előd Egyed-Zsigmond, and Leopold Ghemmogne Fossi. "Bots and Gender Profiling using a Multi-layer Architecture." In: CLEF (Working Notes). September 2019
• Giovanni Ciccone, Arthur Sultan, Léa Laporte, Előd Egyed-Zsigmond, **Alaa Alhamzeh**, and Michael Granitzer. "Stacked gender prediction from tweet texts and images notebook for pan at CLEF 2018." In: Conference and Labs of the Evaluation. 2018, 11p.

## 1.3 THESIS ORGANIZATION

This thesis is structured in seven chapters. In this section, we provide an overview of the content of each chapter.

Chapter 2 presents a conceptual background on our interdisciplinary research topics: Argumentation Theory and FinNLP. Chapter 3 deals with the first research question (*RQ1*) of this thesis, related to the robustness of argument mining model generalization. This chapter consists of two parts, the first

one illustrates our proposed ensemble learning approach towards argument identification task. The second part extends on it, and examines three scenarios: multi-dataset learning, unseen corpora and unseen topics.

Furthermore, Chapter 4 and Chapter 5 focus on earnings conference calls and address the problems of argument modeling (*RQ2*) and argument quality assessment (*RQ3*), respectively. For each of those research questions, we follow graded steps from the state of the art to the derived annotation scheme. We then move to the corresponding annotation study, and report inter-annotator agreement measures. Finally, we study the related computational tasks. Namely, automatic argument identification and argument unit classification in Chapter 4, and classification of argument strength in Chapter 5. In the latter, we further inspect the correlations between our quality dimensions and the declared recommendations of professional analysts.

Chapter 6 considers the application of argument retrieval models to the end of answering comparative questions. Actually, this chapter details on our participation to Touché 2021 which was nominated and published as Best of 2021 Labs. Finally, we summarize our findings, and outline some prospects for future work in Chapter 7.

# 2

# BACKGROUND

In this chapter, we cover the conceptional background of our two lines of research: Argumentation Theory and Financial NLP. Hereby, we inspect the literature terminologies and open challenges, coming to its limitations and research gaps. However, we dedicate further sections of related work along with each of the contribution chapters. In addition, we devote the last section of this chapter to extend on the interplay: argumentation in finance.

## 2.1 INTRODUCTION TO ARGUMENTATION THEORY

Argumentation is a fundamental aspect of human communication, thinking, and decision-making [MO13]. It can be defined as the logical reasoning humans use to come to a conclusion or justify their opinions on a specific topic. The first approaches to study argumentation date back to ancient Greek philosophers in the 6th century B.C.E., and they are known today as the first argumentation theorists. Later on, argumentation acquired more attention from different domains like psychology, communication, linguistics and more recently, computer science. This initiates it as an interdisciplinary research field.

Missimer [Mis95] describes argumentation and its essential components as follows:

> "The objective of argumentation is to *convince an opponent* of a certain *claim*. The claim is a perspective or belief that is justified through logical reasoning. The reasoning is an *inference relation* drawn from *supporting evidence* or reasons towards the claim. If the reasoning is valid, then the claim is a legitimate conclusion of the provided reasons. The manifestation of the application of this process is called an *argument*."
>
> (Missimer, 1995)

Hence, an argument consists of two elementary components: one or more premises leading to one claim [Gov01][1]. Hence, the minimal form of an argument is a single premise connected to a claim, as shown in Figure 2.1.

1 A Premise is also known as the Reason, Justification, or Evidence. A Claim is so-called a Conclusion. We stick to Premise and Claim in this thesis.

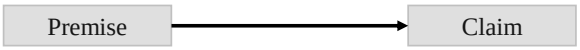



Figure 2.1: The minimal form of an argument, including a claim, a single premise and a consequence relation.

In addition, this definition shows that convincing a reader or a listener by the arguer's opinion is the main goal of argumentation. In fact, argumentation requires a *standpoint* on a topic which supposed to be *controversial* to begin an argumentation.

Grootendorst et al. [GVE04] proposed that increasing (or decreasing) the acceptability of such a standpoint is the objective of argumentation. They also define argumentation as a *verbal* activity, since it is inherently linguistic, either in a spoken or in a written form. And, consequently, as a *social* activity since it implies an interaction with two or more opposing participants:

> " Argumentation is a *verbal*, *social*, and *rational* activity aimed at convincing a reasonable critic of the acceptability of a *standpoint* by putting forward a *constellation of propositions* justifying or refuting the proposition expressed in the standpoint."
>
> (van Eemeren and Grootendoorst, 2004)

Despite that the social feature is more explicit in debates and dialogical communications in general, it is still valid in monological text. This is because, when someone is deliberating on a decision or discussing an issue in an internal monologue, the consideration of costs and benefits is basically a social activity anticipating the reactions of a potential opponent. However, while this definition states that argumentation is a *rational* activity that requires the exchange of reasonable arguments, other facets of arguing such as rhetoric may still play a role [Sau94; BMB10]. Moreover, some parts of the arguments might be implicit, which is known as *enthymematic argumentation*.

Keeping this definition in mind, we can see the need for different argument diagrams and schemes. Each domain expert looks at the argument and the inference structure from a different angle, considering the requirements of the task at hand. Thus, she tries to represent the relations between the premises and claims using a relative scheme. Actually, that leads to one of the main challenges in this domain—the problem of different annotation patterns of available datasets. Consequently, most studies have been concentrating only on one individual sub-task of argument mining, as seen in Figure 2.2:

- Argument Identification: detecting of the argumentative narratives into the text. Thus, classification of a given text into argument or non-argument.

- Argument Components Classification (aka. Argument Unit Classification): it considers the detection of premises and claims.
- Argumentative structure identification: it consists of determining the argument components plus the relations between them [SG14b].

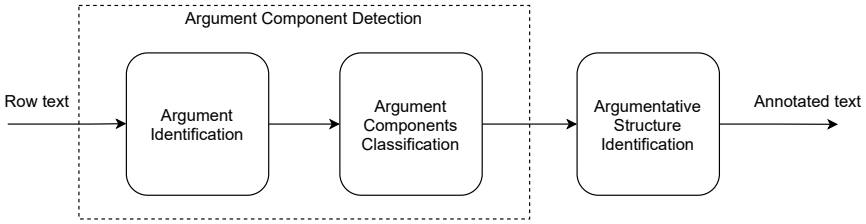



Figure 2.2: Argument mining complete pipeline

In addition, some researchers investigate the argument generation problem (e.g., [Sat+15; SDG20]).

## 2.2 ARGUMENT MODELS

Argumentative discourses are typically presented using one of two directionalities: monologue and dialogue [AK19]. In a monologue, the argument source (e.g, writer) composes and refines his argument flow then delivers it to the target (e.g., reader). This makes the stream of information unidirectional. However, in a dialogue, the argument source and target switch often between participants, making the flow of information bidirectional. Consequently, in a written monological case, the argument source has the opportunity to improve his argumentation strategy before delivering it to the target. For example, optimizing of argument(s) arrangement, using of clear and impressive terms, etc. However, in a dialogical aspect, both sides generate their arguments to be inline with the discussion orientation and the interlocutor's points. However, if this loop of discussion is not valid, this makes the argumentation again as a monological. Nevertheless, in both argumentation directionalities, rhetorical figures (e.g., alliteration and irony) can be used to form a more persuasive speech.

This leads to the three categories of arguments models, shown in Figure 2.3, as proposed by Bentahar et al. [BMB10]:

- Monological models: focus on the internal structure of a single argument (micro-structure).
- Dialogical models: focus on the external 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).

However, those three perspectives on the study of argumentation are closely related [WR03].

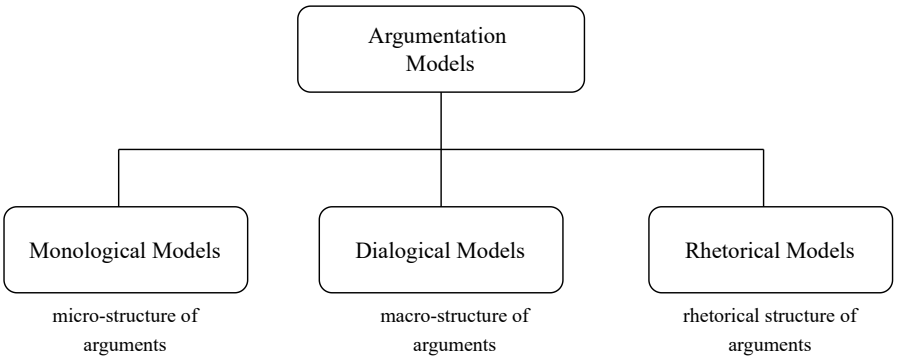



Figure 2.3: Taxonomy of argumentation models adapted from [BMB10]

### 2.2.1 *Argumentation Schemes*

Unlike argumentation structure models, argumentation schemes concern the reasoning type of an argument. In general, three patterns can be observed, inductive, deductive and abductive (aka. defeasible and presumptive). The most popular definition of argumentation schemes and the most quoted one according to Google Scholar [Lum16] is the following:

> ” Argumentation Schemes are forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation“
>
> (Walton et al., 2008)

The first systematic taxonomy of argumentation schemes was proposed in 1962 by Arthur Hastings [Has62]. He defined for each form of scheme a set of critical questions (CQ). The two axes of argument scheme and critical questions work together [WR02]. While the scheme is used to determine the argument's premises and claim, those questions are used to gauge the argument strength, and to reveal its potential weakness.

We illustrate in the following example (which is taken from Walton et al. [WRM08]) how the analysis of argumentation scheme is done in daily arguments:

> *“Helen and Bob are hiking along a trail in Banff, and Bob points out some tracks along the path, saying, “These look like bear tracks, so a bear must have passed along this trail.”* (Walton et al., 2008, p. 9)

In this example, the claim is derived based on the sign (bear tracks). Nonetheless, this claim could be deductively invalid, since the tracks might be from another animal. Therefore, the argument is defeasible and the argumentation scheme of it is known as argument from sign:

ARGUMENT FROM SIGN

**Minor Premise:** A (a finding) is true in this situation.

**Major Premise:** B is generally indicated as true when its sign, A, is true.

**Conclusion (claim):** B is true in this situation.

In this scheme, the argument can be criticized using the two associated questions[2]:
*CQ1*: What is the strength of the correlation of the sign with the event signified?
*CQ2*: Are there other events that would more reliably account for the sign?

Although those questions help to evaluate the argument quality, it is hard to compare two arguments from two schemes [Sta18].

Among other schemes, we also present the scheme of "argument from expert opinion" (sometimes referred to as "Appeal to Expert Opinion") in the following:

ARGUMENT FROM EXPERT OPINION

**Premise:** Source E is an expert in subject domain S containing proposition A.

**Premise:** E asserts that proposition A is true (false).

**Conclusion (claim):** A is true (false).

Despite the natural tendency to respect expert opinion, we can still question this opinion by six critical questions as proposed by Walton [Wal97]:
*CQ1 – Expertise Question:* How credible is E as an expert source?
*CQ2 – Field Question:* Is E an expert in the field that A is in?
*CQ3 – Opinion Question:* What did E assert that implies A?
*CQ4 – Trustworthiness Question:* Is E personally reliable as a source?
*CQ5 – Consistency Question:* Is A consistent with what other experts assert?
*CQ6 – Backup Evidence Question:* Is A's assertion based on evidence?

An example would be:

> "Everybody in Paris says that public transportation is always crowded".

This argument is an appeal to expert opinion, since it is based on the premise that people living in Paris know about its public transportation.

2 Questions are as taken from Walton et al. [WRM08]

Last but not least, we present the scheme of "Argument From Analogy", that deal with argumentation by providing a similar and related evidence.

---

ARGUMENT FROM ANALOGY

---

**Similarity-Premise:** Generally, case C1 is similar to case C2.

**Base-Premise:** A is true (false) in case C1.

**Conclusion (claim):** A is true (false) in case C2.

---

By utilizing our previous example, we can argue about transportation in the capital of Germany:

> "Similarly to Paris, Berlin transportation is often occupied"

Despite the huge insights that Walton schemes enrich the theoretical literature with, he proposed that those models still have to be adapted into "a consistent structure to be useful for formalization and computing" [WR02]. Moreover, one may obviously use multiple inference grounds in one argument, and thus it can follow more than one argumentation scheme [WR03]. The existing list of *96* schemes has been derived from the natural language discourse, and it has been extended by many scholars throughout the decades. Thus, it is uncertain if the current list is ultimately complete. Hence, we believe that argument schemes are still hard to apply in computational argumentation (i.e., automatic argument scheme extraction), and we will not further consider them in our thesis.

### 2.2.2 *Argument Diagramming*

Argument diagramming aims at expressing natural language arguments in a structured visual representation in order to evaluate and analyze them in succeeding steps [Hen00]. Similarly to argument models, the literature differentiates two levels of argument structure:

#### 2.2.2.1 *Micro-level Structures of Arguments*

The micro-level structure of argument is also called argumentation as a product or monological models [HG17]. These models deal with laying bare which argument composition approach the arguer has used. Figure 2.4 exhibits the main recognized approaches. In a basic argument, there is one reason producing a conclusion. In a convergent argument, each of the premises supports the endpoint individually. Whereas, if the premises interdependently serve as a unit to validate a standpoint, this is a linked argument. On the contrary, when tow (or more) conclusions are based on one premise, this is a divergent argument. Finally, in a serial reasoning, each of the premises

supports the other. Apparently, we can find different types of those elementary structures in a complex argument [Hen00].

Hence, previous computational studies allow any sort of composition in the retrieved arguments [SG17a; HG17]. We also adopt this view since it is impossible to know which micro-level structure the speaker will use, or even it is implausible that different speakers (i.e., in our case managers) use the same single strategy of forming their arguments.

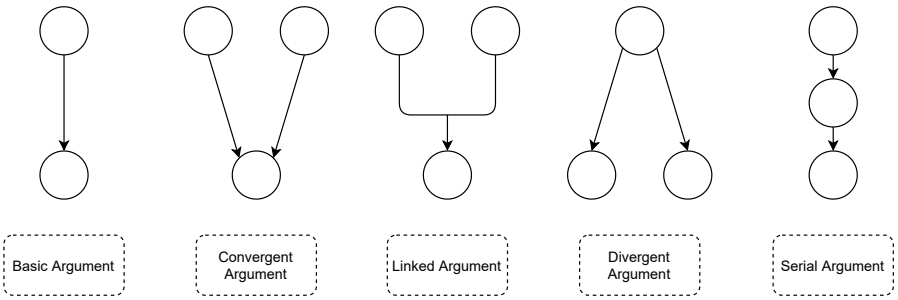



Figure 2.4: Micro-level argument structures

#### 2.2.2.2 *Macro-level Structures of Arguments*

The macro-level structure of the dialogical models of arguments focus on formalizing the connections in a conversation. They consider the speech acts and moves (e.g., assert, challenge) between the two parties (aka. players). Examples of dialogical models are: MacKenzie (1979, 1981) [Mac81] and Amgoud system of modeling dialogues types [AMP00]. However, those models concerned the mathematical modeling of the dialogues.

### 2.2.3 *Toulmin's Model and its Extention*

We conclude the theoretical section of argument models by presenting one widely used conceptual model of argumentation introduced, in the domain of Philosophy of Law [BMB10], by Toulmin [Tou58] [3]. This model is at the micro-level structure and consists of six argument components:

- **Claim** The central point of an argument. All other argument components are used to serve it and increase its truthfulness.
- **Data (Grounds)** It specifies the reasons or facts to establish the foundation of the claim. In other words, it is the counterpart of the common premise [RR05; VE+14].
- **Warrant** To justify why the logical inference from the data to the claim is correct.
- **Backing** The information stands behind the warrant itself. It assures its reliability.

3 Henceforward, we will refer to the updated edition of Toulmin (1958), namely, Toulmin (2003)

- **Qualifier** To what degree of certainty the claim or any other condition should be accepted.
- **Rebuttal** The potential objection situations under which the argument might not hold true. According to Toulmin, it is important that the arguer shows the awareness of possible counter arguments.

The structure of Toulmin's model and the implicit relation between those components can be seen in Figure 2.5 symbolized by arrows and lines. We can agree that by having a data (premise) supporting the claim, we have the minimal form of an argument (cf. Figure 2.1). The warrant can be illustrated as the expressed relation from the premise to the claim. Those components are essential to build an argument, while other parts are supplementary units.

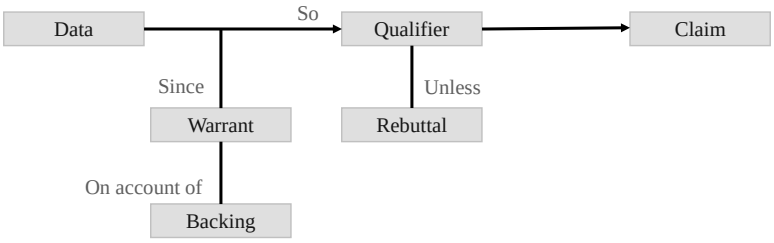



Figure 2.5: Original Toulmin's model of an argument [Tou03]

In his book *The Uses of Argument* [Tou03], Toulmin gave the following example:

> "*Petersen is a Swede;*
> *A Swede is certainly not a Roman Catholic;*
> *So, certainly, Petersen is not a Roman Catholic.*"

The representation of this argument by Toulmin's model can be seen in Figure 2.6.

Despite the fact that this model sets the normative view of a sound argument [Sta18], it has several obstacles for applying it to computational argumentation. First, having all those six parts in one argument is not common in ordinary argumentation. For example, according to [Tou03; VEGSH96], the warrant is almost never declared in daily arguments. Consequently, the same applies for backing. Second, the distinction between the argument components (e.g., Data, Warrant, and Backing) is often indistinct and vague in practice [Hit03; Fre11]. Third, this model allows only a single rebuttal (attacking argument component), without the possibility to defeat that potential counterargument, which is a common strategy in argumentation [Sta18] to enrich the discussion and prevent any future criticism. Based on this point, Habernal et al. [HG17] proposed a *modified Toulmin model* that contains claim, premise, backing, rebuttal and refutation. By that, they reached an inter-annotator agreement

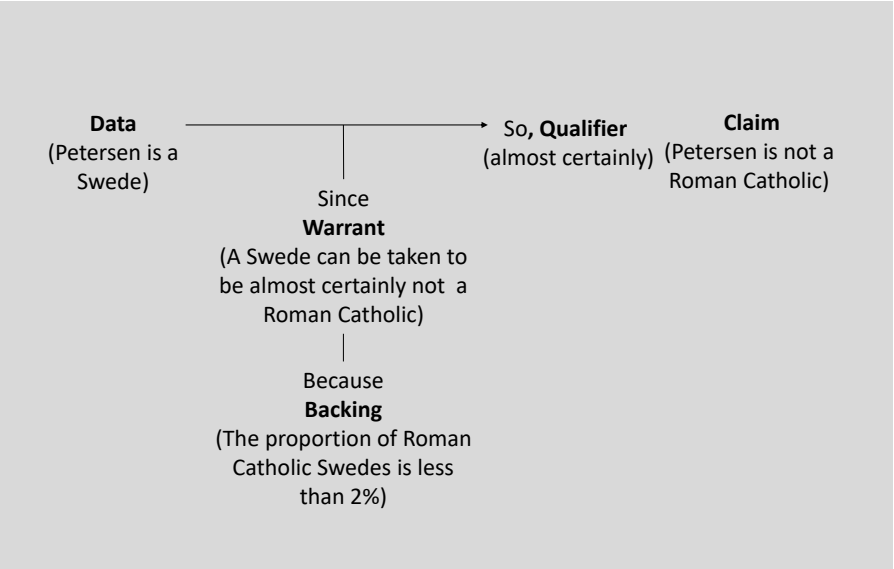



Figure 2.6: Example of Toulmin's model representation of argument structure [Tou03]

of Krippendorff's [Kri04] $\alpha_U$ = 0.48 for labeling 340 documents of web discourse.

Nevertheless, the questions Toulmin defines in his argument components can be used to reveal the weakness points of an argument. For instance, asking for discussing the opponent's view through a "rebuttal" element, empowers the argument against opposing opinions. Thus, this model can be also seen as a tool for evaluating the strength of arguments [SC06].

## 2.3 DESCRIPTION OF THE USED ARGUMENT MINING CORPORA

In our work, we use three publicly available corpora:

The **Student Essays corpus**: contains 402 Essays about various controversial topics. This data has been introduced by Stab et al. [SG14a]. The annotation covers three argument components, namely, 'major claim', 'claim', and 'premise'. Moreover, it presents the support/attack relations between them. Hence, it was used in several argument mining tasks. The dataset also includes one file called 'prompts' which describes the question behind each essay. We consider this 'prompt' as the topic of the essay. They reported Krippendorff's $\alpha_U$ [Kri04] of 0.72 for argument components and Krippendorff's $\alpha$ [Kri80] of 0.81 for argumentative relations.

The **User-generated Web Discourse corpus** is a smaller dataset that contains 340 documents about 6 controversial topics in education, such as mainstreaming and prayer in schools. The document may refer to an article, blog post, comment, or forum posts. In other words, this is a noisy, unrestricted, and less formalized dataset. The manual annotation has been done by [HG17] using a modified version of Toulmin's model [Tou03] with an agreement of Krippendorff's [Kri04] $\alpha_U$ = 0.48 as we have aforementioned in Section 2.2.3.

Table 2.1: Class distributions for all used datasets

| Dataset | #Premise | #Claim | #Non-arg | #Topics |
|---|---|---|---|---|
| StudentEssays | 3510 | 1949 | 1358 | 372 |
| WebDiscourse | 830 | 195 | 411 | 6 |
| IBM | 1291 | 1392 | 0 | 33 |

The **IBM corpus** [Aha+14] consists of, 2683 manually annotated argument components derived from Wikipedia articles on 33 controversial topics. It contains 1392 labeled claims and 1291 labeled evidence for 350 distinct claims in 12 different topics. In other words, there are only 1291 evidences derived from only 12 topics, while there are 1042 claims unsupported by evidence derived from 21 different topics. They achieved an average Kappa agreement [LK77] of 0.39 and 0.4 for claim and evidence confirmations, respectively. This dataset does not include a "Non-argument" label, so we could not use it for the argument identification task. Instead, we used it only for experiments on argument unit classification.

Table 2.1 shows the class distributions for the three datasets. Moreover, different samples of those datasets are expressed in Table 2.2. We can clearly observe that they do not share the same characteristics, like the text length and organization. This makes it more challenging to design a model that generalizes well over them.

Table 2.2: Text examples from the different datasets

| Student Essays | IBM article | Web Discourse |
|---|---|---|
| "First of all, through cooperation, children can learn about interpersonal skills which are significant in the future life of all students. What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. On the other hand, the significance of competition is that how to become more excellence to gain the victory. Hence it is always said that competition makes the society more effective." | "Exposure to violent video games causes at least a temporary increase in aggression and this exposure correlates with aggression in the real world. The most recent large scale meta-analysis- examining *130 studies* with over *130,000 subjects* worldwide– concluded that exposure to violent video games causes both short term and long term aggression in players." | "I think it is a very loving thing, a good and decent thing to send children to a private school! " |

## 2.4 ARGUMENT QUALITY

The question of "what makes a good argument?", can be sometimes suspected with the question "what makes a convincing argument?". However, the latter question is not a product of only a good valid argument but to some external factors too. Aristotle, in his *On Rhetoric* (ca. 350 B.C.E./ translated 2006 [Ken+06]), defined four technical tools of persuasion:

- **Ethos** The level of credibility that the audience has towards the arguer.
- **Logos** The argument has to be logic and reasonable. Different proposals have been suggested in the literature to examine this quality dimension.
- **Pathos** In contrast to logos, this mean of persuasion focuses on the audience sentiment. Namely, by constructing arguments that induct the right emotions, the arguer may win the persuasion game.
- **Kairos** The correct timing and placing of the argument. Nevertheless, this element gained less attention compared to other modes of persuasion.

While the aforementioned means of persuasion may reflect the quality of an argument, it is hard in practice to achieve a correct estimation of all of them. However, the logos seem to be the most independent of other aspects and external factors [JB06]. It has been traditionally studied from two perspectives. First, *formal logic*, where logical arguments, can be either deductive or inductive. Thus, it focuses on verifying the inference relation between the premises and claims [CCR16]. Second, *informal logic*, which consider, mainly, fallacy theories and relevance-acceptability-sufficiency (RAS) criteria proposed by [JB97].

Aristotle contributed with the first list of *fallacies*, which has been extended over decades (e.g., [Dam12; Hab+18]). However, there is still a dispute about the definition of a fallacious argument [WW19] plus if this list is complete or not. A commonly used fallacies are "does not follow up" and "begging the question". We provide in the following an example about "begging the question" argument:

> *Question: How do you know chocolate is good for you?*
> *Answer: Chocolate is healthy because it's good for you!*

On the one hand, examining the argument against a certain collection of fallacies, does not guarantee that it is a good argument. It is just ensuring if it does not have common pitfalls of argumentation. Thus, we do not further count for fallacy theory in this thesis. On the other hand, the RAS-criteria, inspect if all premises of an argument are "relevant", "acceptable" or undoubted facts, and provide enough or "sufficient" evidence for accepting its claim.

Although, informal logic approaches are less restricted and thus more relevant for evaluating arguments in daily life discourse, than formal logic [Gro96],

they are still challenging for an automatic assessment method. Therefore, there have been quite little attempts for evaluating them. For instance, [SG17b] annotated only the "sufficiency" aspect out of the RAS criteria. Their study concerns 1.029 arguments from Student Essays corpus (cf. Section 2.3). Over that, 66.2% were labeled as sufficient arguments, and the rest 33.8% as insufficient. However, the judgment if a given set of premises is sufficient for a certain claim or conclusion, can be quite subjective.

Moreover, the RAS criteria make binary decisions on the argument, based on *all* of its premises. This makes the annotation more likely to be subjective and biased. For example, if an argument with three premises, in which two are strongly relevant to the claim while the last is not relevant, the *all* condition is unfulfilled and so this argument is logically irrelevant. Rather, we would prefer to point out that this argument is relevant to some level. Hence, we suggest our annotation guidelines with respect to a *scale* of assessment.

To sum up, we believe that there are not enough empirical studies on the RAS applicability to real-life arguments. Grootendorst and van Eemeren [GVE04] justified that by the fact that: "Logicians tend to concentrate exclusively on formalized arguments that lack any direction with how argumentation is conducted in practice".

Likewise, practitioners highlighted that theoretical quality dimensions are hard to assess in reality [HG16; Wac+17a]. Therefore, the research of computational argument assessment draws its line, with more practical solutions that do not follow always the normative definitions of philosophical argumentation. Nevertheless, [Wac+17a] was able to demonstrate the relation between the theoretical and practical views of arguments.

To the best of our knowledge, the most recent (and only) survey of existent methods was presented by Wachsmuth et al. [Wac+17b]. They categorize argument quality in three main dimensions (Cogency, Reasonableness, and Effectiveness) as shown in Figure 2.7.

We extend on computational argument quality assessment in Chapter 5 (cf. Section 5.1.1).

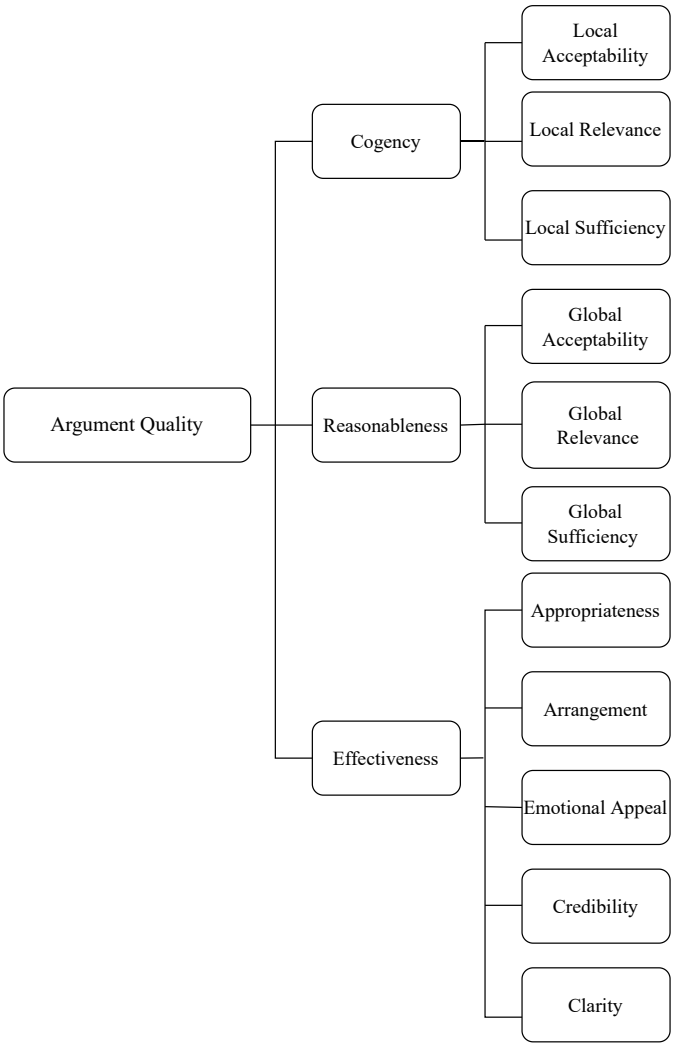



Figure 2.7: A taxonomy of theory-based argument quality dimensions as proposed by Wachsmuth et al. [Wac+17b]

## 2.5 FINANCIAL NATURAL LANGUAGE PROCESSING (FINNLP)

Understanding of financial text has recently gained momentum, which established the Financial Natural Language Processing (FinNLP) as an interdisciplinary field of research [CHC20b; Gup+20].

This section is intended to provide the conceptual background of the stock market domain through two parts. First, we overview the different sources of textual financial information. Second, we discuss the main challenges of stock market prediction and related issues.

### 2.5.1 *Information Sources*

There are multiple ways where financial information is shared with (or between) the public. We can mainly distinguish the following sources.

#### 2.5.1.1 *Insiders*

Before any information becomes well-known to the public, it goes through two stages. Assuming that a fact has been established at time ($t^e$), it is now known by few insiders in the company. However, they are requested by law to keep this kind of information confidential. During this period, this is called *inside information*. At the time $t^p$, company managements release this information to the public. For instance, through an earnings report, which could be two or three months after $t^e$. Once it becomes *public information*, it spreads by the news and social media till it finally turns into well-known information at a time $t^w$ [CHC21b].

Considering this process, the opinions of company insiders are obviously the most critical when analyzing financial instruments. Table 2.3 exhibits the main methods of sharing information with the public. On top of the list, we have the Forms 10-K,10-Q and 8-K. Those three forms are required by the supervision agency, which is the United States Securities and Exchange Commission, also known as SEC [4], and every public company is mandated to fill and publish them.

However, despite the fact that those quarter or annual forms are full of numbers, the count of words is still much bigger than the count of numbers in their contents. Thus, the question remains, how to automatically leverage and convert those words into actionable data that can recommend what to do? Hence, some studies look at particular parts of those forms like the *Management's Discussion and Analysis of Financial Condition and Results of Operations*–commonly known as the *MD&A* section, because it exposes a significantly lower average of similarity across the years in comparision to other sections [CMN20]. Similarly, Zheng et al. [Zhe+19] proposed Doc2EDAG, a financial event extraction method from 8-K reports, in Chinese. They retrieve five event types: equity repurchases, equity freezes, overweight equity, underweight equity and equity pledges.

Table 2.3: The main information sources from stock market insiders

| Source | Explanation |
|---|---|
| Form 10-K | A detailed annual report about the company business, financial conditions, and operations. |
| Form 10-Q | A quarterly report, similar to the annual form 10-K. However, the information here is generally less detailed, and less audited. |
| Form 8-K | A broad form used to notify investors or shareholders about events that occur between 10-K and 10-Q filings. |
| Earnings conference calls | Quarterly organized conferences where a public company discusses the financial results of the last quarter and make estimations about the next one. |
| Speeches or interviews | Managers may be invited to share their view on the industry or their particular companies. |

4 https://www.sec.gov/

In addition to regulatory filings, interviews, public speeches, and any sort of direct communication with the company management provide valuable cues for investors. Among them, earnings conference calls (ECCs) are considered as one of the largest catalysts for variations in stock prices. This conference call takes place typically after the earnings release, where the company shares qualitative and quantitative disclosures about market risk and its profit/loss for the reported period. However, the earnings release, as other official reports, are hard to understand by individual or inexpert investors. Whereas, the earnings conference call offers an understandable alternative (in plain language) and more detailed discussions.

Therefore, ECCs attracted many researchers to investigate on its two sections: the executives' presentation and the Question and Answer (Q&A) session. Prior studies showed that the public management guidance during an earnings call composes a critical input to analysts' forecasting models [Hut05; CTW06].

However, the study of earnings calls transcripts has been mostly addressed by sentiment and semantic features. Yet, some studies considered also vocal features extracted from the call audio recording [Li+20; QY19].

Keith et al. [KS19] 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. They found that ECCs are moderately predictive of post-call analysts' decisions.

In addition, many studies (e.g., [MPR11; Pri+12; ma+20]) found that the question-answer portions of earnings calls have more illustrative and instructive power than the document as a whole. Moreover, given that company executives cannot predict analysts' questions with a complete certainty, executives' responses tend to be more spontaneous and unscripted than in the presentation section[Chi20].

Hence, in our work, we focus only on Q&A sessions especially that it implies also the interaction with the analysts, who we also inspect existent associations between the change of their post-call recommendations and the managers' quality of arguments. To sum up this section, we investigate only on the arguments stated in the answers of company's representatives to the questions of professional analysts.

#### 2.5.1.2 *Professionals*

Market professionals are individuals or organizations that engage in financial market activities such as trading, investment management, and market analysis. Examples include stockbrokers, portfolio managers, investment bankers, and financial analysts.

Although different sources can be investigated similar to the earnings calls (e.g., Form 10-K, Form 10-Q, Form 8-K), Chen et al. [CHC21b] claim that professional analysts' recommendations are generally updated after the earnings

calls. This could be due to the fact, that analysts can get answers and require explanations about their specific points of interest. In other words, analysts either alter or maintain their recommendations based on the information they learned during the call. This means that an analyst would announce a new recommendation only if it does not match the previous one.

Figure 2.8 explains the recommendation scale between one and five, where:

1. **Strong Buy**: also known as "buy" or "on the recommended list", where analysts advise purchasing a specific asset.
2. **Buy**: also known as outperform. It is used when a stock is expected to do slightly better than the market return.
3. **Hold**: when a company is in line with the market and performing at the same pace as comparable companies, a hold recommendation is given.
4. **Sell**: also expressed as underperform/underweight/moderate sell. Meaning that a particular stock is expected to perform slightly worse than the overall stock market return.
5. **Strong Sell**: it's a recommendation to sell an equity or to liquidate an asset.

In the market language, Sell is known as Bearish, while Buy is known as Bullish. Similarly, in some studies, Bearish and Bullish are used as classes of sentiment (i.e., Negative and Positive respectively). In this case, Hold is marked as Neutral.

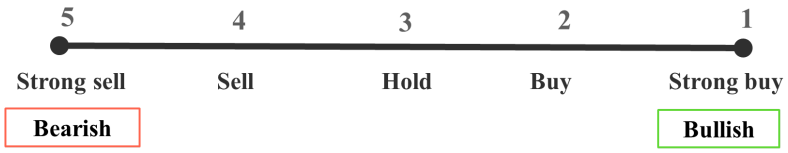



Figure 2.8: Analysts scale of recommendations towards a stock investment

Beside recommendations, analysts also announce their estimations of the share price in the upcoming period. This is known as analyst's price target. The company is performing well when it reaches or outperforms the price target, and vice versa. That explains calling buy as outperform and sell as underperform, as we have seen earlier.

Obviously, those recommendations and price targets formulate an important source of knowledge for investors. An investor can be influenced by professionals opinions, trusting the experience and complete overview they have on a specific stock and on the whole market and economic conditions. Having millions of investors acting towards a stock in the same day (even minute), lead to the price movement, based on the law of supply and demand.
Figure 2.9 simulates the loop of investment decisions and price movement

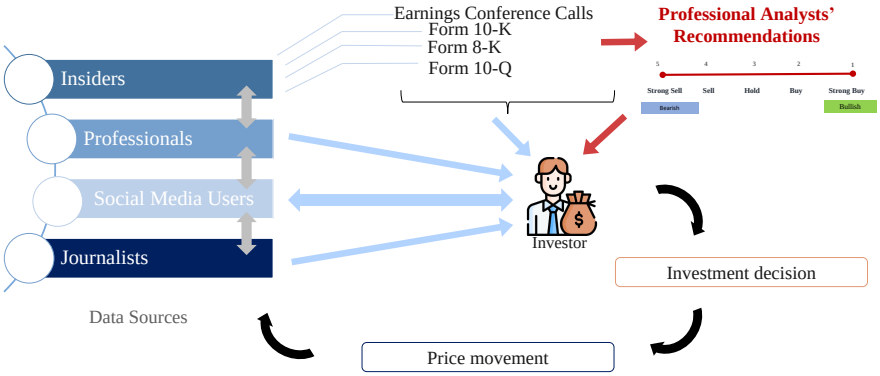



Figure 2.9: Correlation loop between data sources - investment decision and price movement.

based on the updates in information from various data sources, as well as professional analysts' recommendations.

The literature reported many attempts to understand and predict analysts' recommendations, mainly from ECCs as we have explored in 2.5.1.1.

#### 2.5.1.3 *Social Media Users*

Social media platforms have turned into an important source of information for stock market analysis. It can be used by investors to identify trends and inform their investment decisions, as well as by companies to have insights about public opinion and to gauge the effectiveness of their marketing and communication strategies.

One of the most popular platforms, where users express their financial opinions and share market news, is *StockTwits* [5]. Figure 2.10 exhibits an example in which the direct explanation of the post may require some missing words. In general, the automatic understanding of numerals in financial context is still challenging [Che+18]. Moreover, no reason behind this "bullish" opinion, in the example, is given. Moreover, a user may use a URL, a histogram or any image as an evidence for the opinion. Generally speaking, applying argument mining on social media platforms is not always feasible.

From another point of view, anyone can be a social media user: an insider, a professional analyst, an amateur investor or even an organization. Therefore, formulating an opinion on social media must consider the opinion holder herself, since that determines her influence power. In addition, it has to consider the publishing time and the expected validity period of this post [CHC21b].

Some studies look at social media as a material to analyze the differences and association in opinions between the crowd and the experts. For example,

5 https://stocktwits.com/

[EM15] demonstrated the correlation between "wisdom of the crowd" and "analyst opinions" through Granger causality statistical test [Maz15].

Similarly, Chen et al. [CHC21a] aimed at capturing expert-like rationales from social media platforms without the requirement of the annotated data. Zong et al. [ZRH20] found that skilled forecasters express justifications in a more complex language.

Nevertheless, many studies simplify the task to a general sentiment analysis on social media. For instance, [OCA13] collected StockTwits data of six stocks (Google, Apple, Amazon, Goldman Sachsm, Standard and Poors's 500 Index, and IBM) for the period of June 2010 – October 2012. They found that this microblogging data is not sufficient for predicting stock market variables such as returns and volatility.

Guan et al. [GLC22] measured Twitter market sentiment, that is associated with COVID-19 pandemic. More specifically, they use this sentiment to predict the market performance around the *March 2020 stock market crash*. They observed that digital sectors remained resilient comparing to other industries.

All in all, social media has always been one way to gain insights into public sentiment and perception of a particular stock or company, which can in turn influence the stock price.

#### 2.5.1.4 *Journalists*

Journalists may point out the market, industry, or economy changes that convey an important signal for the equity market. They may also interview company management or ask questions during earnings calls.

In most of the cases, business journalists report latest events and facts that consecutively impact the market. For instance, Equifax's stock declined more than 15% following the news of its data breach scandal [ma+20]. Another example would be the Coca-Cola shares, which drop $5 billion after Cristiano Ronaldo's gesture to drink water instead of Coke at the press.

However, while investigating news articles, we have to consider, at least, the two following points:

- Time window: When dealing with news, it is important to decide to which time extent should we still consider an article. This is known as the observation period. Previous studies typically test on a set of


Figure 2.10: Example of a user opinion shared on StockTwits (a social media platform for finance and investement).

different periods (e.g., 1 day, 3 days, 7 days and 15 days). Likewise, in our previous work [Alh+21c], we analyzed the impact of online news (collected from Wall Street Journal, The Washington Post and similar), and StockTwits on the price movement. We used a hybrid approach which consists of sentiment and event-based features as well as the price information for different observation and prediction time windows.

- Opinion holder: In contrary to other data-sources, journalists rarely express their own opinions. A journalist may summarize multiple analysts opinions and social media trends in one article. Thus, extracting the opinion holder requires additional attention in this type of documents [KH06]. In addition, considering the same fact, it could be hard to apply argument mining on news articles since there is no need to argue, rather to transfer opinions and report information. Alternatively, an editorial can be a good source of argumentation since it carries the opinions of the editorial board with the main goal of persuasion. In this regard, Alkhatib et al. [AK+16a] defined six argumentation strategies in 300 news editorials.

### 2.5.2 *Challenges in Financial Document Understanding – NLP and Beyond*

#### 2.5.2.1 *Industry-Specific Terms*

In fact, different industries have their own unique terms and jargon. We started this work by an initial idea that we want to cover companies from various sectors. For example, AbbVie from BioTech industry, and American Airlines from Consumer Discretionary sector. However, soon enough, we realized that understanding related documents lack the knowledge of their specific terms. Hence, we decide to study companies that belong to the Information Technologies sector, since our annotators are mainly computer science students.

Moreover, Loughran and McDonald [LM11] showed that the same word may have positive or negative sentiment based on the discipline. In their large sample of 10-Ks during 1994 to 2008, they found that almost three-fourths (73.8%) of the words identified as negative by Harvard dictionary are words typically not considered negative in financial contexts. For instance, words like taxes, liabilities, and cancer lead to misclassifications in the Harvard list. Similarly, [CHC18; KL14] found that some neutral words in general sentiment dictionary should be considered as the bullish/bearish words. Nevertheless, positive sentiment does not imply a certain bullish market sentiment [CHC20a]. These studies show that even for a simple NLP task like sentiment analysis, we still need to investigate and examine different types of resources. Consequently, there is still an open piece of research related to FinNLP.

#### 2.5.2.2 *The Feasibility of a Good Stock Market Prediction*

Stock market prediction has been always a challenging task as it depends on various factors and is positioned at the interplay of linguistics, machine

learning and behavioral economics [Nas+14]. Traditionally, it has been viewed as a time series prediction problem [ma+20; XC18]. Meaning, exploring trading patterns in the historical market data to forecast future prices (regression problem) or price movement (binary classification problem with up or down). Another research branch considers the external information outside the market data. For example, news, social media, geopolitical status, events or any other related form of content. Therefore, we can classify two main approaches to predict a particular stock value [BPQ20]:

- Technical Analysis: This approach involves analyzing historical price and volume data to identify patterns and trends that may indicate future market movements. Some popular examples are the open/close prices of the stock, the traded volume, but also more complex indicators such as the Relative Strength Index (RSI)[6] and the Consumer Price Index (CPI) [7].
- Fundamental Analysis: This approach involves analyzing a company's financial and economic fundamentals, such as its revenue, earnings, and assets, to determine its intrinsic value and predict its future performance. It also includes the analysis of textual data like news articles and company reports.

In a recent systematic review, Bustos et al. [BPQ20] compared the number of studies with respect to the used approach, as seen in Figure 2.11. They suggest that combining technical indicators along with textual data improves the overall accuracy, as we can observe in Figure 2.12.

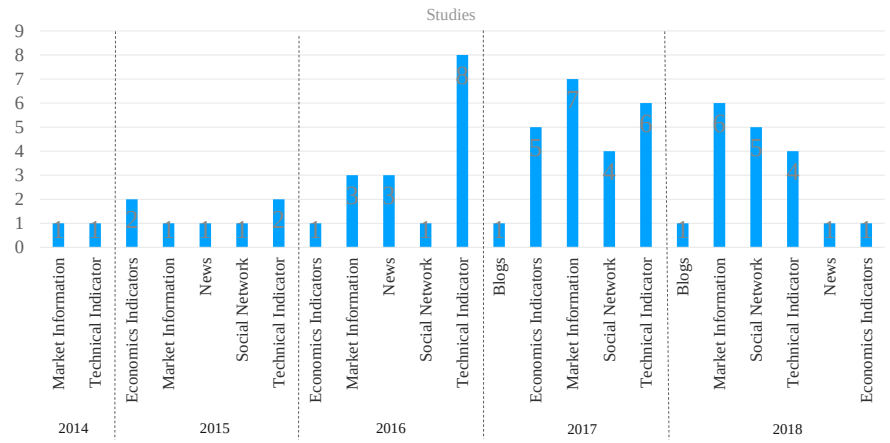



Figure 2.11: Number of published articles per year grouped by data input type, adopted from Bustos et al. [BPQ20]

It is also important to highlight that when targeting stock market prediction, the model accuracy is relatively small. In the review of Bustos et al. [BPQ20]

6 Relative Strength Index (RSI) definition: https://www.investopedia.com/terms/r/rsi.asp
7 Consumer Price Index (CPI) definition: https://www.investopedia.com/terms/c/consumerpriceindex.asp

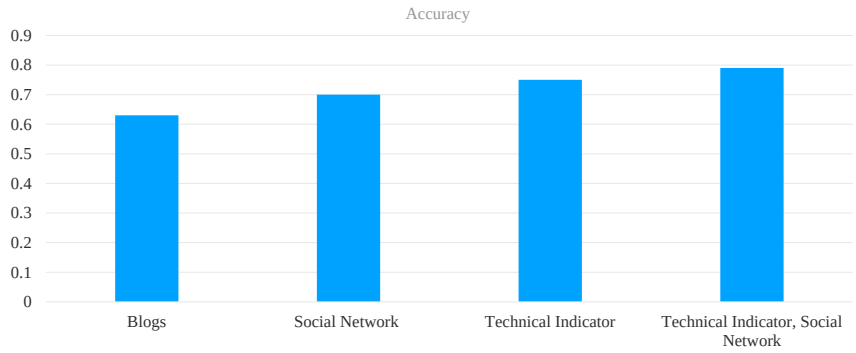



Figure 2.12: Accuracy by input type, adopted from Bustos et al. [BPQ20]

(cf. Figure 2.12), few studies performed close to 80% on their particular datasets. This is normally justified by the random walk theory.

The concept of random walk theory in stock prices was first introduced by the economist Eugene Fama in his PhD thesis *"Random Walks in Stock Market Prices"* which was published in 1995 [Fam95]. In his thesis, Fama argued that stock prices follow a random walk, meaning that they are continuously and unpredictably changing, making it impossible to *consistently* predict future price movements. Moreover, this theory implies that all investors have access to the same information and that the market is efficient, meaning that all stocks are always priced correctly and that there are no undervalued or overvalued stocks. Fama referred to this concept as Efficient Market Hypothesis (EMH). Therefore, he suggested that it would be impossible to consistently beat the market by picking stocks or timing trades. Nevertheless, the real stock market may not always behave in a completely random way, but it's widely accepted that there are random elements that make future stock prices difficult to predict, and that past performance is not a reliable indicator of future performance.

Despite the wide acceptability of Fama's Efficient Market Hypothesis (EMH) and random walk theory, they are still based on assumptions that may not always hold true in the real world. For example, the EMH assumes that all investors have access to the same information and that markets are perfectly efficient, which may not always be the case. Moreover, researchers are still able to define different goals rather than the exact price prediction itself. For instance, developing new investment strategies or risk management techniques that take into account the unpredictability of stock prices. In addition, looking at alternative data sources, such as social media, or news, and use advanced techniques like machine learning to try to extract insights that can be used to predict stock prices (e.g., [Din+15; Alh+21c; Hu+18]), or simply, identify potential opportunities for investors, even if it is difficult to predict the exact price movements.

Therefore, stock market prediction continues to be an active area of research, not only in academia but also in the financial industry.

## 2.6 ARGUMENTATION IN FINANCE

Argumentation in financial domain has been addressed mainly in communication studies in the literature [Pal17; Hur11; Est+10]. Recently, [Paz+19] 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. However, Stab [Sta18] showed that sentiment features are not adequate for support/attack classification task.

Fishcheva et al. [Fis+22] studied the argumentation in *Russian* language and aimed at generating premises for a given claim in the economic domain using ruBERT and ruGPT-3.

On the other hand, there are huge efforts in the FinNLP domain, presented by Chen et al. [CHC21b]. However, most of their work is towards the *Chinese* language (and market) while we consider mainly the English language with respect to S&P 100. Furthermore, they have also organized a series of FinNum tasks that tackle the numerical understanding with respect to the financial text properties. The challenge of 2021, namely, FinNum-3 [8] considers the classification of *in-claim* and *out-of-claim* numerals in the manager's speech during the ECCs[Che+22]. 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. Figure 2.13 shows an example of the data, in which one sentence has two different labels of numerals (in and out of claim). Hence, 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 [ALEZ22].

Based on those studies and on our own experiments on different types of text (e.g., news and StockTwits), we found that earnings conference calls 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 characters count, and people tend to express their opinions and views more than structuring them in a sort of premises and claims. For example, according to our analysis on StockTwits, different posts are only claims with no premises (cf. Figure 2.10).

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

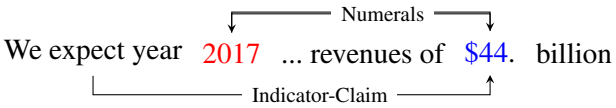



Figure 2.13: An example of the train dataset [CHC20c] including in-claim and out-of-claim numerals.

Therefore, we build henceforth on the transcripts of earnings conference calls, with the aim of analyzing management's persuasion and justification processes through argumentation. Before introducing that in Chapters 4 and 5, and for organization reasons, we inspect argument mining models and domain generalization in the following chapter.

# 3

# DOMAIN GENERALIZATION IN ARGUMENT MINING

The main challenge of the argument mining research field is its variance over domains and topics. The model falls short in shifted domain settings. Therefore, the search for a domain-agnostic model is a point of interest for many researches.

However, to date, the number of studies that address the robustness of the model generalization, rather than on the test split, is still relatively small. Moreover, a unique definition of such a model is missing. On the one hand, most of the works suggest cross-domain models with the mean of integrating multiple heterogeneous datasets in the training process (e.g., [WMS20]). On the other hand, other works define cross-domain as an out-of-distribution testing, where the training and testing datasets (and hence distributions) are related but not the same (e.g., [MS16]). [1]

In this chapter, we aim at filling this gap between computational argumentation and applied machine learning with regard to the model generalization.

Therefore, we survey the related work in Section 3.1. Then, we move to our proposed method based on ensemble learning in Section 3.2. We further apply model selection experiments in Section 3.3. Using this method, and for each of the learned tasks—argument identification and argument unit classification, we address three real-world scenarios concerning the model robustness over multiple datasets, different domains and topics.

Consequently, we first compare single-dataset learning (SDL) with multi-dataset learning (MDL) in Section 3.4. Second, we examine the model generalization over a completely unseen dataset through our cross-domain experiments in Section 3.5. Third, we study the effect of sample and topic sizes on the model performance by means of cross-topic experiments in Section 3.6. We finally summarize this chapter and discuss the future directions in Section 3.7.

## 3.1 RELATED WORK

The goal of domain generalization is to learn a model from one or several different but related domains (i.e., diverse training datasets) that will generalize well to unseen testing domains [Wan+22]. Several research fields are closely

1 In NLP, a domain typically indicates some coherent type of corpus, that is determined by the given dataset [PVN11].

related to domain generalization, including but not limited to: transfer learning, multi-task learning, ensemble learning and zero-shot learning.

Based on that, we define the ***model robustness*** as its performance stability over new variants comparing to the training stage. Mainly, robustness over new data distribution (e.g., [MS16]) or different model runs (e.g., [MML20]).

In computational argumentation literature, and in the last few years, different scholars adopted transfer learning as the solution for cross-domain issue. Among others, Liga et al. [LP20] aimed at implementing a model that discriminates evidence related to different argumentation schemes. They used three different pre-trained transformers (BERT [Dev+18], DistilBERT[San+19] and RoBERTa [Liu+19]) to generate multiple sentence embeddings rather than fine-tuning the models. The resulting embeddings have been used as input for: a support vector machine (SVM), and logistic regression classifiers. Their results show that even with a small amount of data, classifiers trained on sentence embeddings extracted from pre-trained transformers can achieve good scores with respect to the state-of-the-art.

On the other hand, Wambsganss et al. [WMS20] proposed an approach for argument identification using BERT implementation with ten fine-tuning layers and an additional single hidden layer, on multiple corpora. Motivated by this work, we use a fine-tuned DistilBERT as part of our model architecture, and we compare our results with theirs.

Besides transformers, adversarial learning has also been selected to test the model robustness over shifted data distribution by providing deceptive input. Indeed, the assumption behind adversarial learning in NLP, is that by generating variant samples across several domains, deep networks are resistant not only to heterogeneous texts but also to linguistic bias and noise [Zha+20a]. Mayer et al. [May+20] addressed this approach to argument mining (AM), by testing BERT robustness against adversarial examples for argument identification task. Their findings prove BERT efficiency, yet shows that it is not fully invulnerable to simple input perturbations.

Over time, cross-domain argument mining became a must. However, it has been mainly studied in a multi-dataset manner. For example, in the work of Ajjour et al. [Ajj+17], they extend the argument unit segmentation task to investigate the robustness of the model while testing on three different corpora; the essays corpus [SG14a], the editorials corpus [AK+16a], and the web discourse corpus [HG17]. Their proposed argument unit segmentation system is based on a neural network model incorporating features at the word-level for both in-domain and cross-domain settings. Their results show that structural and semantic features are the most effective in segmenting argument units across domains, whereas semantic features are best at identifying the boundaries of argumentative units within one domain. We apply a similar testing scenario in our experiments for both argumentative sentence detection and argument component classification tasks. However, in their study, features are extracted at the word level whereas, we tackle the sentence level classification for our experiments within and cross domains.

In addition, [BAA19] proposed the ArguWeb, a cross-domain argument mining framework based on convolutional neural network (CNN). This framework is designed to first extract the argumentative text, then to classify its units. They used both student essays [SG14a], and the web discourse [HG17] corpora. In terms of cross-domain, the model was trained on one corpus and tested on the other, or on the combined one. Furthermore, they compare the performance of their CNN approach with classical machine learning models, namely, SVM and Näive Bayes [Web10]. Their investigation shows challenging results, since none of the models outperforms the others in all the cases. Hence, we design our method using the concept of ensemble learning to combine the power of both traditional and deep learning models, as we will discuss further in Section 3.2.

In a normal machine learning model training, the dataset is divided into train and test splits. Such that, we evaluate and report the performance of the model on the unseen test split. Yet, this generalization (to the test split) could be limited to data that follow the same distribution which the model has already been trained on. In other words, the model memorized it rather than generalized over it. This issue has been studied by [EHV21] and they conclude that quantifying train/test overlap is crucial for assessing real world applicability of machine learning in NLP tasks, especially when the training data is not large enough.

According to [XM12], the key issue is that the algorithm training error provides an optimistically biased estimation, especially when the number of training samples is small. Therefore, many methods have been suggested to prevent this deviation from the empirical measurements.

Mccoy et al. [MML20] investigated whether the linguistic generalization behavior of a given neural architecture is consistent across multiple instances (i.e., runs) of that architecture. They found that models that differ only in their initial weights and the order of training examples can vary substantially in out-of-distribution linguistic generalization. Therefore, we always consider the average of *five different runs* along all our model generalization experiments in Sections 3.4, 3.5, and 3.6.

## 3.2 ENSEMBLE LEARNING APPROACH FOR ARGUMENT IDENTIFICATION

Ensemble learning is a machine learning research area where different models (i.e., learners) are trained to solve the same problem and combined to get better results [SR18]. The fundamental hypothesis behind it, is that when different models are correctly combined, the ensemble model tends to outperform each of the individual models in terms of accuracy and robustness [LPH07].

This concept of ensemble learning usually comes to the scene with weak learners, such that the overall model is highly improved (e.g., [Gou+19; Cic+18]). In our particular case, we aim to combine the outputs of a classical machine learning model and a transfer learning one, where each individual model provides a considerable performance. Our goal, therefore, is to benefit

from all the features a classical machine learning model uses, and the contextualized knowledge a deep learning model reveals, aiming to improve the performance and stability of the model. To the best of our knowledge, this promising concept has never been used in argumentation tasks.

We employ our experiments on the *argument identification* task, since it is the cornerstone of a complete argument mining pipeline (cf. Figure 2.2). Therefore, our problem is a binary classification task at the sentence-level. To this end, we use only Student Essays and Web Discourse corpora, given that IBM does not imply the "Non-argument" label (see Section 2.3).

### 3.2.1 *Classical Machine Learning Model - SVM*

In terms of the first base model, we consider training a classical machine learning model. This model should be able to capture and learn textual features and patterns that identify argumentative sections of a text. Motivated by the works of [SG14b; SG17a; Moe+07], we defined a set of structural, lexical, and syntactic features in addition to discourse markers as shown in Table 3.1.

The structural features reflect the building of the sentence and its position in the document. For instance, *tokens count* or length of the sentence exploit the fact that premises tend to be longer than other sentences, which can therefore contribute to the argument identification process. Likewise, *question mark ending* indicates that a sentence ending with a question mark is more likely to be a claim, and eventually an argument.

In terms of lexical features, we found that *unigrams and bigrams of Part of Speech (PoS)* tags are very useful to capture the PoS patterns that are frequently observed in argument components. Moreover, *named entity recognition* is a subtask of information retrieval that locates the named entities in unstructured text such as person names, organizations, quantities and time expressions. Such entities are usually used when stating a granted fact, reporting some incidents, or formulating a conclusion (i.e., in an argument component). Therefore, we take into account how many named entities appeared in the sentence as one feature to our model.

Furthermore, syntactic and grammatical features play an essential role for argument identification. In particular, the *depth of parse tree*, the *verbal features* and *count of sub-clauses* which clearly reflect the sentence complexity. This is important since an evidence (premise) tends to appear in a complex sentence structure with more than one sub-clause (in Section 3.3.2, we will see the value of this kind of features on the argument unit classification in particular).

As far as we were able to find out in relevant literature [SG14b], only the tense of the main verb of the sentence has been used to distinguish between claims and premises. However, the tense of the other verbs of the sentence is also helpful to make this identification more accurate. Indeed, sentences including several verbs in the past tense tend to be premises, whereas, the presence of many modal verbs and verbs in the present tense makes the sentence more likely to be a claim.

Table 3.1: The textual features used for argument detection (our newly added features are marked with '*')

| | **Features** | **Explanation** |
|---|---|---|
| **Structural features** | sentence position [SG14b] | indicates the index of the sentence in the document. |
| | tokens count [SG14b; Moe+07] | indicates the count of tokens (words) in the sentence. |
| | question mark ending [SG14b] | boolean feature. |
| | punctuation marks count [SG14b] | indicates how many punctuation marks are there in the sentence. |
| **Lexical features** | 1-3 gram BoW [SG14b; Moe+07] | unigrams, bigrams and trigram Bag of Words features. |
| | 1-2 gram PoS * | unigram and bigram of Part of Speech features. |
| | named entity recognition * | count of the present named entities in the sentence. |
| **Syntactic features** | parse tree depth [SG14b; Moe+07] | indicates the depth of the sentence's parse tree. |
| | sub-clauses count [SG14b; Moe+07] | indicates how many sub-clauses are in the sentence. |
| | verbal features * | counts of [modal, present, past, base form] verbs in the sentence |
| **Discourse markers** | keywords count [SG14b; Moe+07] | number of existing keywords ('actually', 'because', etc.). |
| | numbers count * | indicates how many numbers are there in the sentence. |

Last but not least, we believe that the discourse markers present a direct indicator for argumentative text. For example, the terms: 'consequently' and 'conclude that' are often followed by a claim, while the terms: 'for instance' and 'first of all', are mostly followed by a premise. Hence, we use a set of 286 discourse markers presented by Knott et al. [KD97] to generate the *keywords count* feature that reinforces argumentative text detection. In addition, since statistics are generally used to support a claim, the existence of statistical numbers in a sentence (*numbers count* feature) makes it more likely to be identified as an argument.

We decide to feed those features into SVM classifier. This choice is justified by the fact that SVM performs effectively on small datasets and in high dimensional spaces. In addition, this model provided us experimentally with the best results.

### 3.2.2 *Transfer Learning Model (DistilBERT- based)*

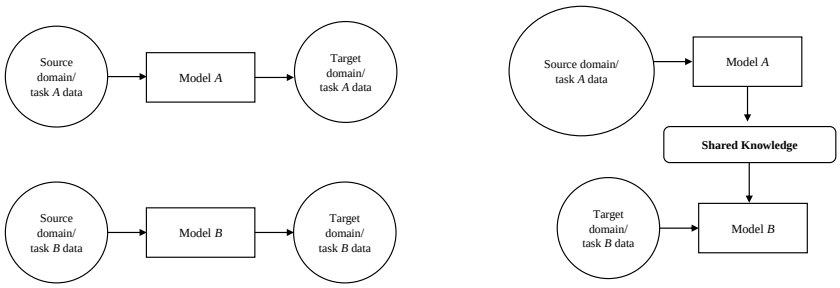
In a traditional machine learning model, there is always an assumption that the training and testing data follow the same distribution and serve the same task (see Figure 3.1a). However, in reality, and in particular for NLP tasks,

this is hard to comply. It is almost impossible to cover all conceivable tasks with annotated data, for all possible domains of text[2], and all the world's languages. To mitigate this problem termed *data scarcity* or so-called *low-resource scenario*, researchers have been working on transfer learning.

Transfer learning seeks freedom from those constraints and searches for mechanisms to adapt models trained on a given dataset to solve slightly different data. It goes towards the search of domain, task, or corpus agnostic models [PY09]. In principle, it aims to apply previously acquired knowledge from one source task (or domain) to a different target one, considering that source and target tasks (or domains) may be the same or may be different but related. This means there is a shared portion of knowledge, which does not need to be learned from scratch, as shown in Figure 3.1b.

The well known example of that, is language representation transfer models, where representations are learned on a large collection of text (e.g., Wikipedia and BooksCorpus for BERT) during the pre-training phase. Then these representations are adjusted to a particular downstream task through fine-tuning phase. This procedure is known as *pre-train then fine-tune* paradigm.

So why is this useful in the argument mining domain? First, common knowledge about the language is obviously appreciable. Second, transfer learning can solve or at least help to solve one of the biggest challenges in the argument mining field, the lack of labeled datasets. Third, even available datasets are often of small size and very domain and task dependent. They may follow different annotations, argument schemes, and various feature spaces. This means that in each potential application for argument mining, we need argument experts to label a significant amount of data for the task at hand, which is definitely an expensive work in terms of time and human-effort. Hence, transfer learning will help us to fine-tune pre-trained knowledge of a large language model (LLM) to serve AM problem.




(a) Traditional machine learning. (b) Transfer learning.

Figure 3.1: The difference between (a) the traditional machine learning setup and (b) the transfer learning scenario.

Among many existent transformers, the Bidirectional Encoder Representations from Transformers (BERT) [Dev+18] has gained a lot of attention. It

2 For instance, political debates, scientific writing, social media, etc.

achieved the state-of-the-art results in several NLP tasks [WMS20; Rei+19; NK19; Cas+20].

For our particular task, we performed different experiments using many BERT-like models (BERT base, RoBERTa-base, DistilRoBERTa, DistilBERT)[3] and achieved very similar results. Hence, we finally decided to use the distilled version of BERT: DistilBERT, given that it is $40\%$ less in size with a relevant in-line performance and a faster training/testing time [San+19].


Figure 3.2: Transfer learning model architecture

In contrast to static word embeddings, the authors of BERT proposed to fine-tune all of its encoder layers. To achieve that, the input has to be tokenized in a BERT-compatible format. This applies, naturally, to its distilled version DistilBERT.

Fig. 3.2 describes the adopted pipeline to perform the text classification using DistilBERT. The first block is the Tokenizer that takes care of all the input requirements: (1) It transforms the sentence's words into an array of DistilBERT tokens. (2) It adds the special starting token ([CLS] token). (3) It adds the necessary padding to have a unique size for all sentences (we set $128$ as a maximum length). The second block is the DistilBERT fine-tuned model, that outputs mainly a vector of length of $768$ (default length). Our mission now is to adapt the output of this pre-trained model to our specific task. We achieve this by incorporating DistilBERT with one additional output layer. This is similar to the way the original BERT paper [Dev+18] suggests fine-tuning it to down stream tasks, such that a minimal number of parameters need to be learned from scratch. Our output linear layer produces a vector of size $2$. The index of the maximum value in this vector represents the predicted class id: argument or non argument. We fine-tune the model for $3$ epochs, using AdamW [LH19] as an optimizer and Cross Entropy for the loss calculation.

### 3.2.3 *Overall Model (SVM + DistilBERT)*

At this step, we have two models based on two completely different approaches. One is based on textual features, while the other is based on a transformer based neural network's ability of language understanding. Since they are two heterogeneous learning models, we chose to use the **stacking** ensemble method to combine their predictions.

Fig. 3.3 presents the stacked model architecture, consisting of two main components: 1. the base models, that include the trained transformer based model (DistilBERT) and the trained SVM model in parallel, and 2. the meta-model, that will learn from the outputs of the two models to produce the final prediction of a sentence. In order to have an array of independent features for

3 In all of our experiments, we used transformers from Huggingface - https://huggingface.co/ .

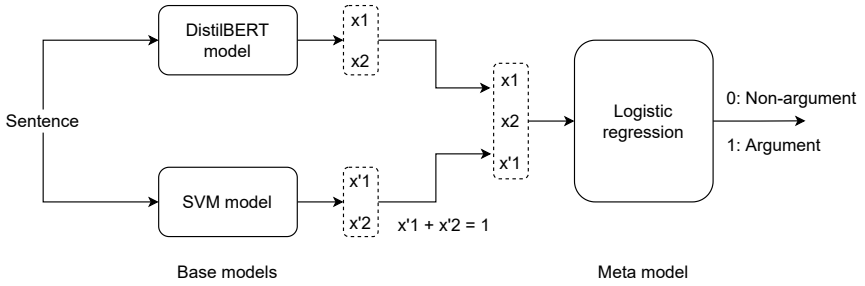



Figure 3.3: Stacked model architecture for argument identification task

the meta-model, and since SVM outputs two probabilities $x'1$ and $x'2$ (i.e., $x'1 + x'2 = 1$), we consider only $x'1$. Whereas, $x1$ and $x2$ are two independent raw logits so both of them are considered. Given that we are dealing with a binary classification problem where the input features are independent, logistic regression serves well as a meta-model to accomplish the task. For the training/testing steps, we first split the combined dataset into $75\%$ training and $25\%$ for the overall testing. This testing data remains unseen for all the models, and it is used only for the final validation of the overall model. The base models are trained on the $75\%$ training data. The training data of the meta-model is prepared by $5$-folds cross validation of the two base models. In each fold, the out-of-fold predictions are used as a part of the training data for the meta-model.

### 3.2.4 *Evaluation*

In this section, we discuss the performance of each of the individual learners apart and the final stacked model. In addition, we state some comparison with the most recent previous work [WMS20] tackling the same problem of argument identification on the same datasets. We will further discuss the results shown in Table 3.2, 3.3 and Table 3.4.

In terms of SVM model, we can see that it works very well on the Essays corpus with an accuracy of $90.95\%$ and F1-score of $83.75\%$. Yet, it seems less efficient on the Web Discourse corpus, where the transfer learning model provides better measurements. This can be interpreted by the formal structure of Student Essays compared to Web Discourse, as we elaborated in Section 2.3. On the merged corpora, SVM achieved an accuracy of $85.42\%$, using the textual features that learn a set of patterns in argument identification. In some cases, SVM fails to classify an argumentative sentence as an argument due to the absence of language understanding capabilities this task needs rather than the representative features[4]. These limitations might be handled by understanding the meaning of the sentence using the characteristics of transfer learning through the pre-trained DistilBERT model. Evidently, there are other

4 Here is an example (from Essays dataset) of an argument sentence that SVM fails to identify while DistilBERT succeeds: *"Personally, I think both government and common people should have the responsibility for the environment, but we need to analyze some specific situations."*

cases where the contrary happens – SVM classifies correctly and DistilBert model fails [5]. Hence, we have decided to combine these models.

Table 3.2: Evaluation on **Student Essays** corpus

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **SVM** | 0.9095 | 0.8730 | 0.8116 | 0.8375 |
| **DistilBERT** | 0.8727 | 0.8016 | 0.7477 | 0.7697 |
| **Stacking model** | 0.9162 | 0.8890 | 0.8195 | 0.8483 |

As we can see in the normalized confusion matrices (Fig. 3.4), SVM model reaches a higher percentage than DistilBERT-based model in terms of True Positive (TP) whereas the latter performs better than SVM for True Negative (TN). Therefore, the stacked model is getting the most out of both of them in terms of TN and TP, and thus it records a better classification accuracy, precision, and recall as shown in Table 3.4.

Table 3.3: Evaluation on **Web Discourse** corpus

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **SVM** | 0.7437 | 0.7051 | 0.5882 | 0.5874 |
| **DistilBERT** | 0.7799 | 0.7718 | 0.6484 | 0.6655 |
| **Stacking model** | 0.7855 | 0.7449 | 0.6958 | 0.7113 |

In a recent work [WMS20], the authors implemented a BERT-based transfer model on different corpora including the two datasets we have used. Our stacked model overcomes theirs on the Student Essays achieving an accuracy of 91.62% and F1-score of 84.83% compared to their accuracy of 80.00% and F1-score of 85.19%. On the Web Discourse corpus, we have similar accuracy values (78.5% to 80.00%) while on the level of the combined model, our approach achieved better performance even though they have investigated on more training corpora. [6]

Table 3.4: Evaluation on the merged Corpora (**Student Essays and Web Discourse**)

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **SVM** | 0.8542 | 0.8037 | 0.7012 | 0.7331 |
| **DistilBERT** | 0.8587 | 0.7887 | 0.7529 | 0.7683 |
| **Stacking model** | 0.8780 | 0.8326 | 0.7659 | 0.7921 |

5 Here is an example: *"nowadays, there is a prevailing opinion that human needs for farmland, housing, and industry are more important than saving land for endangered animals."*

6 They used *AraucariaDB* and *Blog-comments* corpora. The former has received additional annotations and modifications over the years as part of the AIF-DB. As of today, the dataset does not include the original text anymore, and we could not use it. Regarding *Blog-comments* corpus, they mentioned that it is part of the Wikipedia Blog Comments introduced by [BR11]. This corpus was rarely adopted in the literature, and we did not find it.

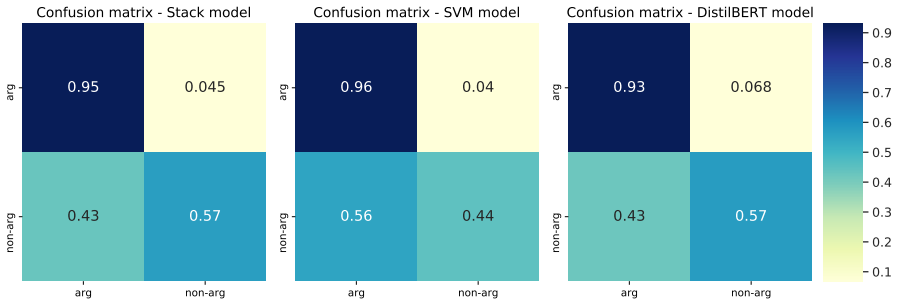



Figure 3.4: Normalized confusion matrices

Furthermore, we suggest that the idea of combining two different approaches is not only about the improvement of results, but also a step forward for the model's interpretability. On the one hand, deep learning models (transformers in our case) reduce the task of feature engineering. Yet, it is difficult for humans to easily fully understand their behavior. On the other hand, the direct feature engineering involved in classical machine learning makes those models more interpretable and easier to customize.

## 3.3 MODEL SELECTION

Before we move to the generalization experiments, we extend our work to the argument unit classification task, using the three corpora (cf. Section 2.3). Furthermore, we apply deep feature analysis and model selection, motivated by the work of [SG17a] to set up the best model configuration for each of the two addressed AM tasks, namely argument identification and argument unit classification.

In fact, including more features in the training can be problematic since it can increase space and computational time complexity. It can also introduce some noise according to unexpected value changes. These shortcomings are known as the curse of dimensionality. Even though SVM is able to deal with high dimensionality, we believe that removing redundant features can reduce overfitting and improve hold out performance.

The main solution for dimensionality reduction is feature selection, where different methods can be applied. In our work, we have first applied a simple *filter method* that is based on variance threshold such that we can figure out any features that do not vary widely between the three classes. We achieve that by visualizing the distribution's histogram of each feature. This helps us to see if a feature is important and improves the performance, or if it has a redundant effect (or even no effect) on the final output. We present two examples in the following:

Figure 3.5 suggests that sentence position in the input paragraph correlates positively with premise sentences. In particular, with the positions 1 to 5. This means that a sentence that is stated earlier in the paragraph is more likely to be a premise than a claim or non-argument. We have also observed that the value of position zero is very frequent since in Web Discourse (WD) and

IBM datasets, we do not have long paragraphs like in the Student Essays (SE), rather it may be only one sentence (and hence have the position 0).

Similarly, Figure 3.6 reflects the distribution of the punctuation marks over the three classes. We obviously can see that non-argument text tends to have more punctuation marks than argumentative text. Also, in terms of premise/claim classification, sentences with more than seven punctuation marks are only premises.

Both "sentence position" and "number of punctuation marks" are part of our structural features, which proved to be very essential in our model selection process. We identify the best performing model by conducting a *feature ablation tests*.

Consequently, in order to determine the best configuration for our stacked model, we apply at this step a kind of *wrapper method* that iterates through different combinations of features and performs a model retrain on each. For this model assessment, we adopt the accuracy as well as the weighted average metrics of precision, recall and F1-score. That's because our data is imbalanced, and our priority is to detect and classify the argumentative sentences correctly, which correspond to the larger class.

The feature combination which produces the best model performance metrics for each AM task is selected. Since the effect of different groups of features will be on the SVM performance in the first place, and subsequently on the stacking model that combines SVM with DistilBERT predictions, we report in this section, both SVM and stacked model results for the different settings. Moreover, in order to ensure more statistically significant testing, we have conducted for every set of features 5 runs over 5 different seeds, and internally 5-fold cross validation. That means for each set of features, the model is tested 25 times. We report the weighted mean and the standard deviation of those runs for each classification task.

### 3.3.1 *Model Selection on Argument Identification Task*

Table 3.5 shows the results of argument identification task using SVM over different groups of features. Our findings suggest that SVM scores the best performance using lexical, structural and syntactical features with a slightly better weighted F1-score of 85.7% than SVM with all features or with lexical,

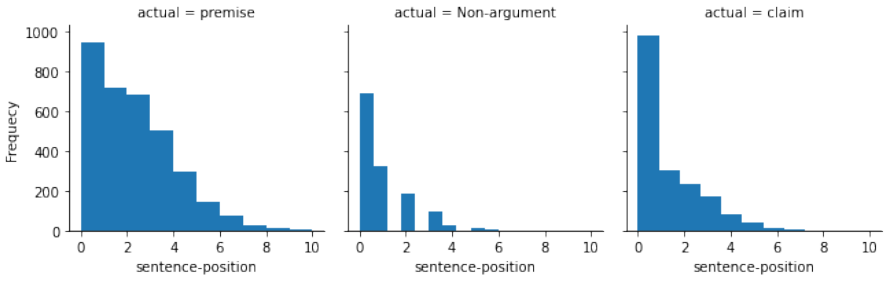



Figure 3.5: Histograms of the *sentence position* feature

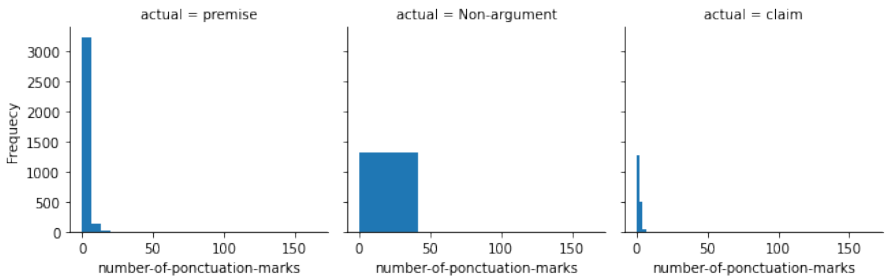



Figure 3.6: Histograms of *number of punctuation marks* feature

structural and discourse markers (W-F1 score = 85.6%) while they all achieve the same accuracy of 86.1%.

Table 3.5: Results of feature analysis on argument identification task using SVM on SE and WD

| | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| lexical | 0.782 | ±0.001 | 0.807 | ±0.001 | 0.794 | ±0.001 | 0.807 | ±0.001 |
| structural | 0.825 | ±0.0 | 0.838 | ±0.0 | 0.831 | ±0.0 | 0.838 | ±0.0 |
| syntactic | 0.617 | ±0.0 | 0.786 | ±0.0 | 0.691 | ±0.0 | 0.786 | ±0.0 |
| discourse markers | 0.617 | ±0.0 | 0.786 | ±0.0 | 0.691 | ±0.0 | 0.786 | ±0.0 |
| lexical, structural | *0.849* | ±0.001 | *0.858* | ±0.001 | *0.853* | ±0.001 | *0.858* | ±0.001 |
| lexical, structural, syntactical | **0.853** | ±0.001 | **0.861** | ±0.001 | **0.857** | ±0.0 | **0.861** | ±0.001 |
| lexical, structural, discourse markers | *0.852* | ±0.0 | *0.861* | ±0.0 | *0.856* | ±0.0 | *0.861* | ±0.0 |
| all features | *0.852* | ±0.0 | *0.861* | ±0.0 | *0.856* | ±0.001 | *0.861* | ±0.0 |

Similarly, Table 3.6 confirms that the combination of structural, lexical and syntactical features achieves the best performance at the level of the stacked model. However, we observe that the scored mean of different settings is similar, especially when considering the structural features. According to the Student t-test [DW13], when structural features are considered, the p-value exceeds 5%. Hence, we cannot claim that including (excluding) some features, except for structural and lexical, makes a huge difference on our model. Nevertheless, we adopt the best performing model which empirically proved to be the one with structural, lexical and syntactical features for argument identification task. Henceforth, we use these settings for the upcoming experiments on this particular task.

### 3.3.2 *Model Selection on Argument Unit Classification*

To train the model on argument unit classification (i.e., premise/claim classification), we transform the feature "*Keywords count*" that indicates the count

Table 3.6: Results of model selection on argument identification using the stacked model on SE and WD

| | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| lexical | 0.830 | ±0.004 | 0.842 | ±0.003 | 0.836 | ±0.003 | 0.842 | ±0.003 |
| structural | 0.851 | ±0.006 | 0.86 | ±0.005 | 0.856 | ±0.006 | 0.860 | ±0.005 |
| syntactic | 0.831 | ±0.006 | 0.843 | ±0.005 | 0.837 | ±0.006 | 0.843 | ±0.005 |
| discourse markers | 0.831 | ±0.007 | 0.843 | ±0.006 | 0.837 | ±0.007 | 0.843 | ±0.006 |
| lexical, structural | *0.862* | ±0.002 | *0.869* | ±0.002 | *0.866* | ±0.001 | *0.869* | ±0.002 |
| lexical, structural, syntactical | **0.863** | ±0.003 | **0.870** | ±0.003 | **0.866** | ±0.003 | **0.870** | ±0.003 |
| lexical, structural, discourse markers | *0.861* | ±0.004 | *0.868* | ±0.004 | *0.865* | ±0.004 | *0.868* | ±0.004 |
| all features | *0.861* | ±0.003 | *0.868* | ±0.002 | *0.865* | ±0.003 | *0.868* | ±0.002 |

of any argument indicator, to two features: "*premise-indicators-count*" and "*claim-indicators-count*".

Furthermore, we also integrate a new dataset: IBM (cf. Table 2.1) and we further employ the model selection experiments as in the previous AM task. Table 3.7 confirms that SVM with all features delivers slightly better results compared to the other sub-combinations of features.

Table 3.7: Results of feature analysis on argument unit classification task using SVM on SE, WD and IBM datasets

| | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| lexical | 0.802 | ±0.001 | 0.803 | ±0.001 | 0.802 | ±0.001 | 0.803 | ±0.001 |
| structural | *0.840* | ±0.0 | *0.841* | ±0.0 | *0.841* | ±0.0 | *0.841* | ±0.0 |
| syntactic | 0.378 | ±0.0 | 0.615 | ±0.0 | 0.468 | ±0.0 | 0.615 | ±0.0 |
| discourse markers | 0.633 | ±0.0 | 0.648 | ±0.0 | 0.640 | ±0.0 | 0.648 | ±0.0 |
| lexical, structural | *0.847* | ±0.001 | *0.848* | ±0.001 | *0.847* | ±0.001 | *0.848* | ±0.001 |
| lexical, structural, syntactical | *0.846* | ±0.0 | *0.846* | ±0.0 | *0.846* | ±0.001 | *0.846* | ±0.0 |
| lexical, structural, discourse markers | *0.846* | ±0.0 | *0.847* | ±0.0 | *0.847* | ±0.001 | *0.847* | ±0.0 |
| all features | **0.848** | ±0.0 | **0.848** | ±0.0 | **0.848** | ±0.0 | **0.848** | ±0.0 |

In terms of the stacked model, beside the semantic conceptual features that DistilBERT learns, we observe that the structural features are the most dominant proprieties that help to discriminate premises from claims in the three used corpora. However, they achieve a slight difference in comparison to their combination with lexical features and to the all features performance, as shown in Table 3.8. This finding is similar to the one by [Ajj+17], which shows that structural and semantic features are the most effective in segmenting argument units across domains.

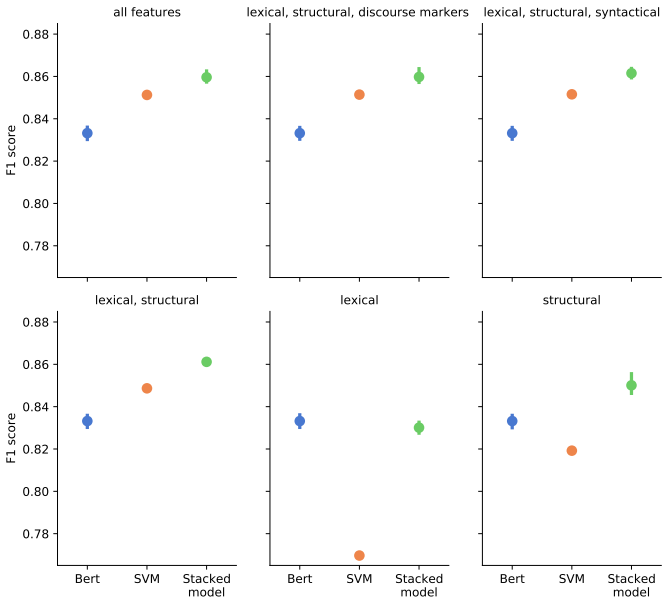



Figure 3.7: Effect of feature selection on the argument identification task

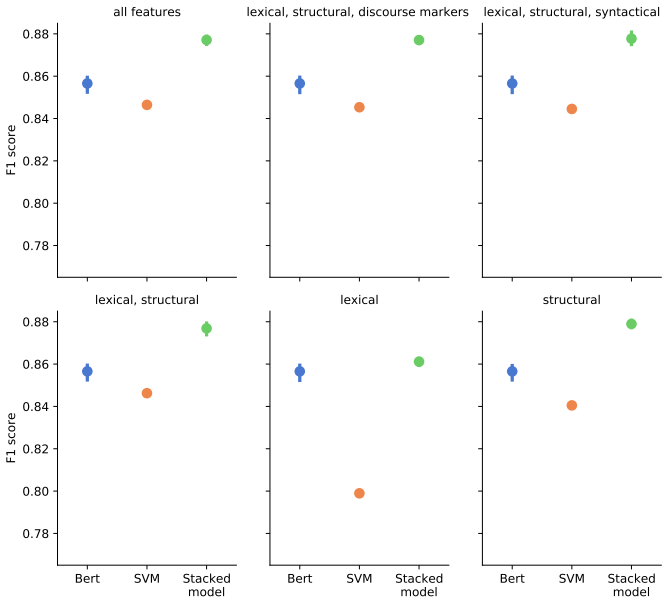



Figure 3.8: Effect of feature selection on argument unit classification task

Table 3.8: Results of model selection on argument unit classification task using the stacked model on SE, WD, and IBM datasets

| | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| lexical | 0.862 | ±0.002 | 0.863 | ±0.002 | 0.862 | ±0.002 | 0.863 | ±0.002 |
| structural | **0.88** | ±0.002 | **0.88** | ±0.002 | **0.88** | ±0.002 | **0.88** | ±0.002 |
| syntactic | 0.857 | ±0.003 | 0.857 | ±0.002 | 0.857 | ±0.002 | 0.857 | ±0.002 |
| discourse markers | 0.858 | ±0.003 | 0.858 | ±0.003 | 0.858 | ±0.002 | 0.858 | ±0.003 |
| lexical, structural | *0.878* | ±0.003 | *0.878* | ±0.003 | *0.878* | ±0.003 | *0.878* | ±0.003 |
| lexical, structural, syntactical | *0.878* | ±0.004 | *0.879* | ±0.004 | *0.879* | ±0.004 | *0.879* | ±0.004 |
| lexical, structural, discourse markers | *0.878* | ±0.002 | *0.878* | ±0.002 | *0.878* | ±0.002 | *0.878* | ±0.002 |
| all features | *0.878* | ±0.002 | *0.878* | ±0.002 | *0.878* | ±0.003 | *0.878* | ±0.002 |

Furthermore, Figures 3.7 and 3.8 report the F1 scores, and standard deviation, for SVM, DistilBERT and the stacked model across the different sets of features. We can observe that the stacked model scores at least the same performance as DistilBERT, and it improves over once the SVM classifier obtains a minimum score of 80% which is verified in most cases.

To sum up, in Section 3.3, we applied an in-depth feature analysis and model selection in two-folds: argument identification and argument unit classification. According to our findings, we ignore, henceforth, the features that lead to minor short-term wins, and we keep only the structural features for argument unit classification, and structural, lexical and syntactical features for argument identification task.

## 3.4 MULTI-DATASET LEARNING

This experiment is intended to determine whether incorporating more datasets in the training step will generate a significant, positive impact on the robustness of the stacked model with respect to the test data, taking into account that our available datasets are relatively small. Consequently, we compare the outcomes of single-dataset learning (SDL) and multi-dataset learning (MDL) approaches.

In the SDL setup, we train and test the model on each dataset individually while in the MDL setup, we train the model on all datasets, but test on individual test splits (20%) of a particular dataset. This methodology allows us to report performance scores on each dataset separately while training our model on a single versus multiple datasets.

We examine our model in these settings for the two trained tasks; argument identification and argument unit classification. However, since IBM has only the labels of argument components, we run the argument identification experiments using WD and SE datasets, whereas we use WD, SE, and IBM for the

argument unit classification experiments. We use for each task its best stacked model configuration conducted in Section 3.3.

Table 3.9: SDL vs. MDL argument identification using the stacked model.

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SDL | SE | SE | **0.918** | ±0.002 | **0.92** | ±0.002 | **0.919** | ±0.002 | **0.92** | ±0.002 |
| | WD | WD | **0.771** | ±0.014 | **0.776** | ±0.011 | **0.773** | ±0.014 | **0.776** | ±0.011 |
| MDL | Merged | SE | 0.877 | ±0.006 | 0.881 | ±0.004 | 0.879 | ±0.004 | 0.881 | ±0.004 |
| | Merged | WD | 0.749 | ±0.011 | 0.765 | ±0.009 | 0.757 | ±0.015 | 0.765 | ±0.009 |

According to Table 3.9 and Table 3.10, we observe an expected drop in the performance for all datasets between the SDL and MDL setups. Yet, our stacked model is still able, in all the cases, to produce reliable accuracy and F1-score. Nevertheless, detecting argumentative text proved to be an intrinsically more generalized task than determining the premises and claims. For example, the variation of F1-score between the two settings, is in the range of [-2%,-4%] for argument identification, while it moves to the range of [-7%,-9%] for argument unit classification task.

These evaluation results also suggest that a single learning is always better when we are sure that our future targeted data follows the same or a very close distribution to the training one. This allows a better capturing of the dataset characteristics. On the other hand, in a multi-dataset approach, merging the datasets may introduce some noise if the model does not have enough samples to weight the particular traits of the tested data. Therefore, despite the fact that available argumentation corpora are small, we cannot merge them to improve the model performance over one of them.

Table 3.10: SDL vs. MDL argument unit classification using the stacked model.

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SDL | SE | SE | **0.825** | ±0.003 | **0.827** | ±0.003 | **0.826** | ±0.003 | **0.827** | ±0.003 |
| | WD | WD | **0.888** | ±0.012 | **0.868** | ±0.01 | **0.878** | ±0.007 | **0.868** | ±0.01 |
| | IBM | IBM | **0.987** | ±0.002 | **0.987** | ±0.002 | **0.987** | ±0.002 | **0.987** | ±0.002 |
| MDL | Merged | SE | 0.736 | ±0.134 | 0.738 | ±0.125 | 0.737 | ±0.127 | 0.738 | ±0.125 |
| | Merged | WD | 0.802 | ±0.026 | 0.796 | ±0.015 | 0.799 | ±0.014 | 0.796 | ±0.015 |
| | Merged | IBM | 0.913 | ±0.006 | 0.895 | ±0.008 | 0.904 | ±0.009 | 0.895 | ±0.008 |

## 3.5 CROSS-DOMAIN SETTINGS: TESTING ON A COMPLETELY UNSEEN DATASET

The hypothesis behind the model generalization in machine learning, is its performance over the test split which stays unseen during the training process. However, this assumption has a couple of caveats based on the fact that we are drawing our test samples identically from the same distribution, and thus we are not biasing ourselves in any way [Wan+22]. Hence, and in order to answer the question: to which extent is our approach independent of the domain and data diversity, we adopt another examination of the model robustness over shifted or cross-domain settings. That is to say, we are testing on a completely new corpus and not only a subset of unseen samples from the same training corpus. Consequently, this approach is also known as *out-of-domain* (OOD) testing. However, it has been referred to as cross-domain in different argument mining studies (e.g., [MS16]). Therefore, we apply our experiments in a hold-out manner. In other words, we keep out in each run one dataset for testing and we train on the remaining ones. We again assay in these experiments our stacked model with only structural features for argument unit classification and with structural, lexical and syntactical features for argument identification task (cf. Section 3.3). We report the weighted mean and standard deviation over 5 different seeds.

The outcomes of cross-domain argument identification and cross-domain argument unit classification are presented in Table 3.11 and Table 3.12, respectively.

Table 3.11: Evaluation of the cross-domain argument identification task.

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Training | Testing | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SE | WD | stacked model | 0.559 | ±0.006 | 0.455 | ±0.013 | 0.502 | ±0.013 | 0.455 | ±0.013 |
| | | DistilBERT | **0.661** | ±0.003 | **0.694** | ±0.003 | **0.677** | ±0.002 | **0.694** | ±0.003 |
| | | [MS16] | | | | | 0.524 | | 0.524 | |
| WD | SE | stacked model | 0.749 | ±0.006 | 0.771 | ±0.012 | 0.760 | ±0.006 | 0.771 | ±0.012 |
| | | DistilBERT | **0.759** | ±0.006 | **0.798** | ±0.005 | **0.778** | ±0.004 | **0.798** | ±0.005 |
| | | [MS16] | | | | | 0.128 | | 0.181 | |

In terms of argument identification task, and based on the empirical evaluation presented in Table 3.11, we observe a satisfactory performance of our stacking model (W-F1 score= 0.76) when training on WD and testing on SE. However, the opposite scenario drastically reduces the performance, where (W-F1 score= 0.502). While those are both better than the results of [MS16] who used a binary statistical classifier with a similar set of our SVM features. DistilBERT is still able to outperform the stacking model in this scenario.

In regard to the argument unit classification (Table 3.12), we observe that training the stacked model on SE plus IBM and testing on WD yields worse

results than training on other datasets (W-F1 score=0.627). However, it is still outperforming DistilBERT when testing on IBM and SE. In fact, the performance of DistilBERT degraded for this task, especially when testing on SE, and it achieves its best performance when testing on WD. That means, for premise/claim classification, we still need the features of SVM that allow our stacked model to overcome transfer learning once the tested corpus implies a formal structure that could be better learned using traditional machine learning. This also interprets the worst case of stacked model (trained on SE, IBM and tested on WD), since WD does not imply such learned features (e.g., sentence position) and by contrary, SVM pulls back the stacked model performance in this testing scenario.

To sum up this section, our results suggest that transferring knowledge across different datasets is more applicable for argument identification task. Comparing to the stacked approach [Alh+21b], DistilBERT is still reaching a higher accuracy when fine-tuned on the same dataset. This means that transfer learning is very efficient for in-domain-generalization, and less efficient for cross or out-of-domain generalization. However, this is even more challenging for argument unit classification where the ensemble learning model shows a better generalizing capability, in most cases, with the power of learning genre-independent presentations of argument units. We further apply cross-topic testing in Section 3.6.

## 3.6 CROSS-TOPIC SETTINGS: TESTING ON COMPLETELY UNSEEN TOPICS

In this section, we further assess the stacked model performance and compare it with DistilBERT, over unseen data, with a finer-grained level of cross-settings referred to as cross-topic. In this experiment, we aim to study whether the model performance over unseen topics will be improved by considering more training topics, or by considering more samples for each training topic. In other words, the analysis will reveal whether the diversity of sampling (a wide range of topics) improves cross-topic performance.

Table 3.12: Evaluation of the cross-domain argument unit classification task.

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Training | Testing | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SE, WD | IBM | stacked model | **0.766** | ±0.015 | **0.610** | ±0.052 | **0.679** | ±0.081 | **0.61** | ±0.052 |
| | | DistilBERT | 0.704 | ±0.028 | 0.550 | ±0.013 | 0.618 | ±0.024 | 0.55 | ±0.013 |
| SE, IBM | WD | stacked model | 0.735 | ±0.08 | 0.546 | ±0.281 | 0.627 | ±0.303 | 0.546 | ±0.281 |
| | | DistilBERT | **0.773** | ±0.008 | **0.805** | ±0.009 | **0.789** | ±0.004 | **0.805** | ±0.009 |
| WD,IBM | SE | stacked model | **0.677** | ±0.013 | **0.675** | ±0.016 | **0.676** | ±0.044 | **0.675** | ±0.016 |
| | | DistilBERT | 0.356 | ±0.128 | 0.586 | ±0.128 | 0.443 | ±0.141 | 0.586 | ±0.128 |

### 3.6.1 *Experimental set-up*

To perform these experiments, we derive a group of new datasets out of the SE, WD, and IBM datasets according to each particular classification task. The number of sentences per topic (|S|) varies across the three datasets. However, we still need to unify the size of data for all tested combinations, as well as unifying the |S| in each. By that, we only analyze the effect of diversity sampling (|T|) on the model generalization to unseen topics.

Accordingly, to perform the cross-topic experiments, we have to satisfy the equation:

$$N = |T| * |S| \tag{3.1}$$

where:

- $N$ is the fixed size of each new dataset.
- |T| is the number of topics (variable).
- |S| is the number of sentences per topic (variable).

To satisfy this equation, the first step is to fix $N$ in a way that we can have multiple pairs of (|S|, |T|). This implies that a higher |T| leads to a lower number of sentences per topic |S|. Three constraints need to be fulfilled:

1. We need to maximize $N$ to have a sufficient size of data.
2. $N$ should allow obtaining different combinations of |T| and |S| with respect to our corpora statistics, which is challenging given how number of sentences per topic varies across the three datasets: For example, |S| in SE corpus varies from 7 to 46, while |S| in WD corpus varies from 76 to 362 sentences given that it has only 6 topics in total.
3. We need to include samples from all three sources (SE, WD, and IBM) to form a mixed dataset.

The second step, after fixing the pairs (|S|, |T|), is to derive a group of new datasets out of the SE, WD, and IBM datasets according to each particular classification task.
For the argument identification task, we fix $N$ to 1200 and the pairs (|S|, |T|) to (4, 300), (6, 200), and (24, 50).
Likewise, for the argument unit classification task, we fix $N$ to 1200 and the pairs (|S|, |T|) to (3, 400), (4, 300), (6, 200), and (24, 50).

The third step is to design the test sets. For generalization purposes, we run the cross-topic over 5 runs (5 seeds) and internally over a 5-fold cross-validation setup. We report the average mean and standard deviation of the weighted precision, recall, F1 score, and accuracy on the testing set. Our 5-fold cross-validation is in terms of topics. In other words, the train set covers 80% of the topics and the remaining unseen topics are in the test set.

### 3.6.2 *Evaluation*

In the following, we present the obtained results in Table 3.13 and Table 3.14 for argument identification and argument unit classification, respectively. For argument identification task, the evaluation results prove that the stacking model performance is consistent over the different sets of topics: W-F1 score averages between 0.810 to 0.893, and the accuracy ranges from 0.813 to 0.895. Similarly, in the unit classification task, the W-F1 averages between 0.801 to 0.858, and accuracy ranges from 0.80 to 0.855.

Table 3.13: Model assessment in cross-topic experiments for argument identification task. S: number of sentences per topic, T: number of Topics

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| S | T | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 4 | 300 | stacked model | **0.892** | ±0.019 | **0.895** | ±0.019 | **0.893** | ±0.021 | **0.895** | ±0.019 |
| | | DistilBERT | 0.765 | ±0.083 | 0.825 | ±0.03 | 0.794 | ±0.052 | 0.825 | ±0.03 |
| 6 | 200 | stacked model | **0.855** | ±0.009 | **0.862** | ±0.008 | **0.858** | ±0.009 | **0.862** | ±0.008 |
| | | DistilBERT | 0.703 | ±0.089 | 0.791 | ±0.021 | 0.744 | ±0.036 | 0.791 | ±0.021 |
| 24 | 50 | stacked model | **0.807** | ±0.029 | **0.813** | ±0.026 | **0.81** | ±0.032 | **0.813** | ±0.026 |
| | | DistilBERT | 0.626 | ±0.09 | 0.775 | ±0.03 | 0.693 | ±0.041 | 0.775 | ±0.03 |

These findings suggest that the ensemble learning stacking approach is outperforming DistilBERT in all the cases with W-F1 score approximately +10% for argument identification and up to +5% for argument unit classification. Moreover, the former reported a lower variance in the standard deviation for almost all tested cases. This is in line with the findings of [MML20] who found that 100 instances of BERT are remarkably consistent in their in-distribution generalization accuracy, while they varied dramatically in their out-of-distribution generalization performance. Therefore, since a BERT-like model (DistilBERT in our case) is less stable to completely unseen data, the stacked approach gets a valuable impact on the model robustness in such out-of-distribution or cross-domain scenarios. Moreover, according to Zhang et al. [Zha+20b], BERT only exploits "plain context-sensitive features" such as character or word embeddings. It poorly deals with incorporating "structured semantic information".

In terms of the impact of |T| and |S|, the weighted F1 score has been improved by increasing the |T| in the training set for the argument identification task. However, the opposite behavior is observed concerning the argument unit classification task: i.e., increasing the |T| decreased the weighted F1 score. We explain this contrast by the influence of the vocabulary employed in each task. In fact, the structure of arguments may differ according to the discussed topic. For instance, we can find more statistical arguments in finance and more logical well-structured arguments in law. Therefore, ensuring distinct and diverse

samples (varying topics during the training process) is important to generalize the learned patterns of argumentative text. However, for the argument unit classification, distinguishing between premise and claim is more related to the grammatical structure of sentences which does not require topic-specific vocabulary. For instance, we can use claim keywords (consequently, in fact, implies) or premise keywords (such as because, moreover, since) to distinguish between the argument components.

Table 3.14: Model assessment in cross-topic experiments for argument unit classification task. S: number of sentences per topic, T: number of Topics

| | | | **W-Precision** | | **W-Recall** | | **W-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| S | T | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 3 | 400 | stacked model | **0.802** | ±0.013 | **0.800** | ±0.014 | **0.801** | ±0.013 | **0.800** | ±0.014 |
| | | DistilBERT | 0.774 | ±0.02 | 0.767 | ±0.023 | 0.770 | ±0.022 | 0.767 | ±0.023 |
| 4 | 300 | stacked model | **0.822** | ±0.03 | **0.82** | ±0.032 | **0.821** | ±0.034 | **0.82** | ±0.032 |
| | | DistilBERT | 0.764 | ±0.032 | 0.766 | ±0.031 | 0.765 | ±0.031 | 0.766 | ±0.031 |
| 6 | 200 | stacked model | **0.825** | ±0.019 | **0.825** | ±0.019 | **0.825** | ±0.02 | **0.825** | ±0.019 |
| | | DistilBERT | 0.789 | ±0.02 | 0.786 | ±0.019 | 0.787 | ±0.019 | 0.786 | ±0.019 |
| 24 | 50 | stacked model | **0.861** | ±0.054 | **0.855** | ±0.055 | **0.858** | ±0.056 | **0.855** | ±0.055 |
| | | DistilBERT | 0.847 | ±0.074 | 0.835 | ±0.079 | 0.841 | ±0.076 | 0.835 | ±0.079 |

In addition, we report the macro-average results of all of our model generalization experiments in Appendix a. However, they still imply the same conclusions in terms of the outperforming model.

## 3.7 CONCLUSION

We addressed in this chapter two main problems of argument mining: argument identification and argument unit classification. Our study is on the sentence-level with a stacked ensemble learning approach. We aimed to detect the essence of argumentative text and to assess the robustness of our model in more realistic scenarios than testing on a subset of the data known as the test split.

While generalization has always been an important research topic in machine learning research, the robustness and generalization of argument mining models are yet not well explored. This is a very urgent task to elevate the research in this field given the two-fold challenges it has: the lack of labeled data, and the domain dependency performance of the existing models. We believe that a formal protocol of testing the model generalization and robustness is an instant need in argumentation domain, since every scientific paper tackles it from only one angle. Most of the works suggest cross-domain models, with the mean of integrating more datasets in the training process.
According to [Wan+22], developing more fair and application-driven evalua-

tion standards, as well as the interpretability of the results, are one of the most challenging open issues in domain generalization.

Therefore, in our work, we defined sets of experiments that infer an empirical evidence on the model performance in real world applications. Based on our comparison of single-dataset learning (SDL) and multi-dataset learning (MDL), we propose that SDL is always recommended when we are confident that the future dataset will be similar to the training one. Furthermore, our findings suggest that knowledge transfer is more applicable for argument identification than argument unit classification in cross-domain (out-of-distribution) setup. In terms of the latter task, the stacked model outperformed DistilBERT when tested on IBM and SE corpora. This indicates that recognizing premise and claim texts is more related to the structure of the sentence. A similar conclusion is reached in our cross-topic experiments on this particular task, where we found that the more |S| (number of sentences per topic) we have for training, the better the stacked model generalizes to unseen topics. However, the sampling diversity (increasing the topic count |T|) was essential for the argument identification task, such that topic-specific vocabulary plays a crucial role.

Since the structure of the sentence made a difference in many of our experiments, we plan to test if providing a transfer learning approach (e.g., DistilBERT) with such features, would outperform the ensemble learning approach based on this enriched knowledge. This research direction is towards the understanding of how transformers indeed work, and how we can develop them [RKR20]. In our future work, we also plan to run joint model experiments where argument identification and argument component classification are in one sequential pipeline. We also plan to investigate more on the segmentation model that predicts the boundaries of the argument and on optimizing the combination of the base models (SVM and DistilBERT).

# 4

# ARGUMENT MINING IN EARNINGS CONFERENCE CALLS

In this chapter, we consider the second research question (*RQ2*) raised in the introduction. As discussed in Chapter 2, particularly in Section 2.6, there is no completely financial corpus annotated with fine-grained argumentation structures despite the existing attempts. On the other hand, our AM model presented in Chapter 3 still needs to be trained on (at least on some of) the target data to be efficient at automatically detecting argumentative text and classifying it into its argument units.

As we have mentioned in Section 2.5.1.1, the ECCs provides a forum for managers to relay company operations to individual and institutional investors. Moreover, it provides the opportunity to respond to professional analysts' questions about the company performance and expected earnings. Different studies found that the discussion during the question answering session is the most informative and influencing part on the market [MPR11; Pri+12]. In their study to predict the stock price movement based on ECCs, Ma et al. [ma+20] found that including the presentation section does not improve the model performance.

Moreover, given that company representatives cannot predict analysts' questions with complete certainty, their answers tend to be more unscripted than in the presentation section [Chi20]. Therefore, in our work, we focus only on Q&A sessions. In other words, we investigate only on the arguments stated in the management's responses to the questions of professional analysts.

To the best of our knowledge, no prior work has been carried out to annotate arguments in earnings calls transcripts. Therefore, the contributions of this chapter are the following:

- First, we propose an annotation scheme, derived from argumentation theory, for modeling arguments in the answers of Q&A sessions of earnings conference calls.
- Second, we present our annotation study and the reliability of the created labels by means of inter-annotator agreement, on 15% of the data, with four annotators.
- Third, we evaluate our data using the argument mining stacking approach introduced in Chapter 3.

- Fourth, we provide our annotated *FinArg* corpus under free license to encourage future research in both computational argumentation and FinNLP domains [Alh+22b][1].

## 4.1 ANNOTATION SCHEME

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. As we have seen in Chapter 2, argument diagramming is a good foundation for modeling argumentation structures since it allows separating and distinguish several arguments in the text [SG14b] through the relation that connects the corresponding components with each other.

Although the Q&A session implies a bidirectional conversation, the assumption that the target of persuasion is changed in this conversation is not valid. That is because the analyst is only asking a question, and in very rare cases following up with a new one to the same person. Hence, the dialectical rules of judgment we have discussed in Section 2.2 are not applicable here. Indeed, the arguments are presented only in the portion of managers' answers and not in the questions. Therefore, we focus on the monological perspective of argumentation models, which is also convenient for developing computational methods [PS13].

Chen et al. [CHC21a] suggested to use Toulmin's model to structure argumentation in analysts' opinions (in analysts' reports). However, as we have detailed in Section 2.2.3, this model has several drawbacks to model the daily life argumentation [HG17; PM09; Fre11]. Therefore, we do not follow this model to structure our data. Instead, we adopt a simpler annotation scheme based on the minimal requirements to form an argument.

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 linked. 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 put forward evidences (premises) that strengthen 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). Instead, in the scope of our work, we are interested in detecting the arguments themselves as independent units.

More details about the data and the annotation study will be given in Section 4.2. We first explain and accentuate the selected scheme of argument structure.

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

**ARGUMENTATION STRUCTURE AND ARGUMENT COMPONENT TYPE**

Similarly to previous works (e.g., [SG14a]), our annotation scheme models the structure of argument as a node-link diagram. Each node represents an argument unit (i.e., a premise or a claim), and each link represents a directed argumentative relation which could be either a support or attack. However, assigning an argumentative type to the argument component could be ambiguous in deep structures. For instance, it is fuzzy to distinguish the between a backing that supports the warrant of a premise (as in Toulmin's model 2.2.3), from the backing that supports the whole argument (as in the modified Toulmin's model [HG17]). Moreover, it is more fuzzy to separate this latter (backing that supports the whole argument) from an ordinary premise component[2].

Therefore, we choose to label any statement supporting the final standpoint of the arguer either in a direct or in-direct way as a premise. In particular, we model the structure of *each argument* using *one-claim-approach* proposed by Cohen [Coh87]. 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 symbolizes the relation, which could be either a support or an attack.

The argumentation literature introduced basically four approaches for assigning the argumentative type to an argument unit:

- First, the *one-claim-approach* we have aforementioned.
- Second, the *multi-label-approach*, where an argument component can have two labels. For example, in a serial argument structure, a claim of one argument is at the same time a premise for another argument [Bea50].
- Third, *level-approach*, which specifies a certain label to each level. For instance, [Gov10] differentiate between a "main claim" and "sub-claim". Likewise, [Dam12] identifies "premise" and "sub-premise".
- Fourth, Stab et al. [SG14a] proposed a *hybrid-approach* that combines the level-approach with one-claim-approach. Thus, in their Student Essays dataset, they distinguish between "major claim" and "claim", but still model each argument in the one-claim-approach.

To sum up, in our annotation scheme, we consider the *one-claim-approach* for structuring every single argument. However, we observed in our data that some speakers elaborate simultaneously on different sides of the controversy, and raise an argument for each side. Therefore, we allow multiple arguments to be labeled in one document. Yet, and similarly to [HG17], we restrained the annotators from creating complex argument hierarchies.

Figure 4.1 represents a sample of our annotation scheme, which implies that we can have diverse types of argument micro-structures (cf. Figure 2.4) in one answer.

2 In their definition[HG17], "the argument should still make sense after removing the backing".

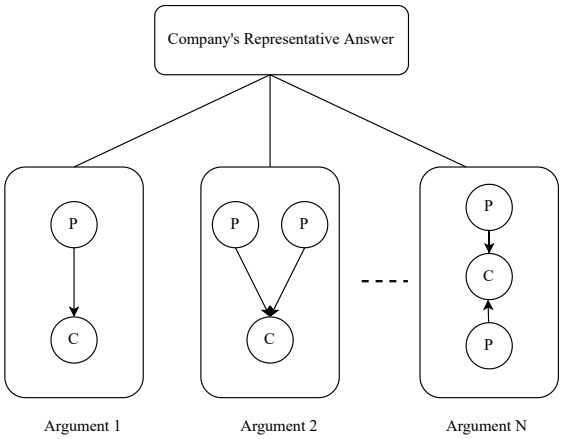



Figure 4.1: Argument annotation scheme (a sample) including argument components and argumentative relations (support/attack) indicated by arrows

Moreover, you can see a real example of the data in Figure 4.2. 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 (marked in italics face), then the speaker starts to argue about his claim ($C_1$) by stating different premises. The annotator labeled every sentence ($P_1$ to $P_3$) as a premise since they all emphasize the stance of the speaker, and build a chain of reasoning leading towards the claim. In this particular example, all those premises belong to the same claim, and they are all marked with a support relation type.

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 ] **P1**. [ But it's very good for us to see the breadth of subscriptions that we offer and that customers are interested in ] **C1** . *It's very large.* [ And if you remember, we quoted the same number a quarter ago and we talked about 150 million ] **P2**

[ 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 ] **P3. [ ... ]**

Figure 4.2: An example fragment of the Apple Q2 2017 earnings call conference transcript—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 $\mathbf{P}_{count}$.

## 4.2 FINARG CORPUS CREATION

The motivation for creating a new corpus is threefold:

- First, we believe that it is time to reason the financial data and to move from shallow linguistic features and opinion mining to the reasons behind it, the analysis of persuasion 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 applications [CHC20b].
- Third, the same challenge applies for argumentation field, where available datasets are often of small size and very domain and task dependent [HG17]. Therefore, our dataset can serve the computational argumentation scholars as well.

### 4.2.1 *Annotation Setup*

**Data**: We downloaded the transcripts of ECCs through a paid membership in the Financial Modeling Prep API [3].
**Annotation tool**: We used the free tool of Label Studio [4] as a visualized annotation tool (see Appendix d).
**Annotators**: We have hired four non-native English-speaking students, but with an excellent level of language. Three of them are computer science students, while the fourth is doing his master in international economics and business.
**Annotation Scope**: Our annotated data covers the quarterly earnings calls of four companies: *Amazon* (AMZN), *Apple* (AAPL), *Microsoft* (MSFT) and *Facebook* (FB), during the period of: *Q1 2015- Q4 2019*. Thus, we have 80 earning call transcripts in all.

As we have discussed in Section 2.5.2.1, the choice of those companies is based on the fact that other industry sectors imply industry-specific terms that are not comprehensive for our annotators. Nevertheless, even with technology companies, they had to search and understand some financial terms in a plain language through websites like "Investopedia" [5] and "The Motley Fool" [6]. This also anticipates that the final trained model may not be efficient for every different type of industry. This is, however, the general rule for any supervised machine learning model.

For each *transcript*, we created a list of all the *speakers*. After having determined the role of each of them *(Analyst, Representative, or an Operator)*, we were able to split the whole text into different documents. Each *document* contains one or two questions asked by a *single analyst* and the corresponding

3 https://site.financialmodelingprep.com/developer
4 https://labelstud.io/
5 https://www.investopedia.com/
6 https://www.fool.com/

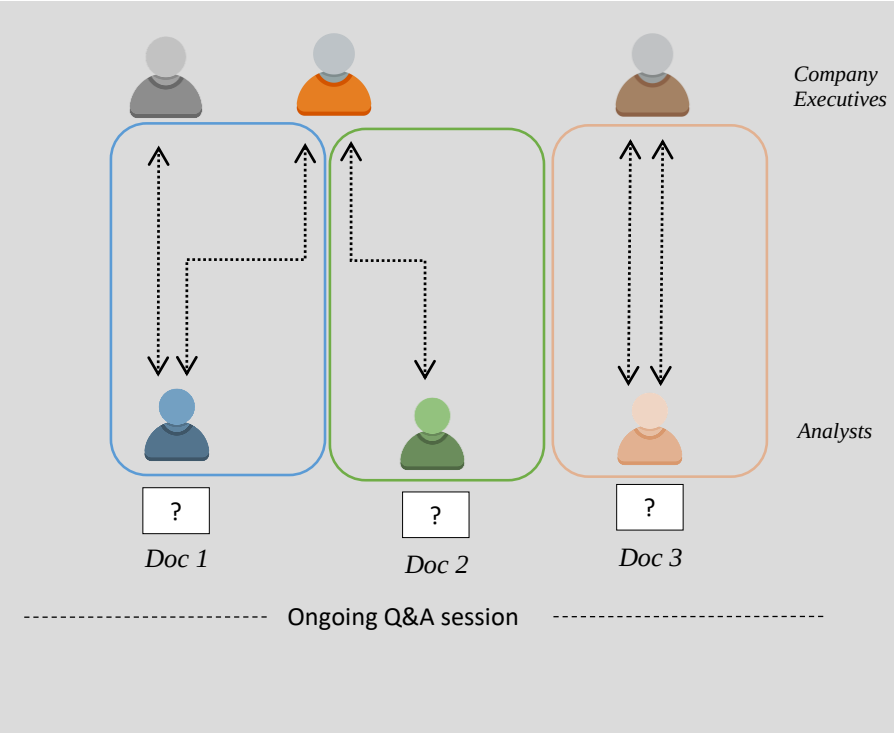



Figure 4.3: Simulation of Q&A session, and the representation of a document in our data

response(s) by the company representatives (see Figure 4.3)[7]. We reformatted these documents following Label Studio guidelines, and imported them to be labeled with argument units and relations, as well as identifying the non-argumentative text.

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 raises two questions, and receives one (combined) or multiple 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 4.1).

### 4.2.2 *Annotation Study*

Our annotation study consists of three stages:

1. Annotation guidelines: We conduct a preliminary study on a set of documents, to define the annotation guidelines, while elaborating with one of our annotators. The detailed guidelines can be seen in Appendix b.

2. Pilot annotations: The goal of this stage was to test the annotation guidelines before a complete corpus is annotated. This was done by

7 Henceforth, we refer as a *document* to this fragment of text: (question-answer) pairs corresponding to one analyst.

training sessions and discussions with the annotators. We got their feedback to refine 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 needs *two to three hours* to be completely annotated. This supports our choice of annotation at the micro-structure level of argument and with the one-claim-approach.

Moreover, to let annotators gain insights into the company's performance over the years, we assign one company to each of our four annotators.

3. The study of Inter Annotator Agreement (IAA): We compute how homogeneous and thus reliable the annotations are (to be explored in Section 4.2.2.2).

#### 4.2.2.1 *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) – without any overlapping between the components. 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. Automatic Speech Recognition (ASR) systems generally do not produce accurate punctuated transcripts [Fu+21]. The quality of the punctuation marks we got in our data is, therefore, based on the system used to generate those transcripts. In our case, we let the annotators segment each span of text based on the context with respect to the splitting rules we have defined earlier.

#### 4.2.2.2 *Inter-Annotator Agreement (IAA)*

To evaluate the reliability of our data, we determine a group of 12 complete earnings calls that represent about *15%* 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 unitized alpha $\alpha_U$ [Kri04] and Krippendorff's $\alpha$ [Kri80] for the argument

components and argument relations annotations, respectively. That is because, to the best of our knowledge, Krippendorff's $\alpha_U$ is the only applicable agreement measure when both the annotation unit boundaries, and the label to be annotated at the same time [AP08].

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.

Nevertheless, it can be tricky to get identical annotations given that the argument component types are strongly related. In other words, the annotation of a premise depends on its connected claim. Therefore, every permutation of two annotators had to meet and discuss their disagreement cases to produce the last validated document *(gold annotations)*.

ANALYZING DISAGREEMENTS As a result, we discovered that the primary source of uncertainty for argument components is due to the missing of unit boundaries, and the connected context that covers multiple sentences. In terms of argument relation, we think that the uncertainty is due to the high ambiguity of argumentation structures, as it was also previously noted by Walton [Wal96]. That is, when diagramming arguments, there are many cases where there is a room for more than one interpretation [Hen00]. Moreover, Stab et al. [SG14a] found that even with a pre-identified argumentative components, there are often multiple valid interpretations of an argumentative relation between them, i.e., it is "[...] hard or even impossible to identify one correct interpretation" [SG14a].

We also asked the annotators to read the entire question to identify the controversial topic before starting with the actual annotation task on the answer paragraph. Although this approach is more time-consuming than a direct identification of argument components and relations, it yields to a more reliable annotated data. Furthermore, understanding of the question will help to assess the quality of arguments, which we will address in Chapter 5.

### 4.2.3 *Creation of the Final Corpus: the FinArg Dataset*

Once the annotation is complete using Label Studio, the output file is a very long JSON[8] document. However, before using this data, we ran some scripts to detect if any annotation errors exist. Most often, a document was classified to have an error because of (at least one of) these three issues: the answer part of the document was not fully annotated (e.g., missing to cover one word), the same piece of text was annotated twice (caused by accidental clicks on Label Studio), or a relation was misdirected (e.g., from the claim to the premise instead of the opposite).

When it is possible, the issue was corrected automatically by code. Otherwise, we ask the corresponding document's annotator to correct that mistake.

8 JavaScript Object Notation.

Thereafter and to increase the usability and reproducibility of our *FinArg* dataset, we arrange the output annotations file with a similar style format to the Student Essays dataset's annotation file [SG14a], since it is simply understandable and probably the most used corpus in computational argumentation.

Hence, the annotation document (file.ann) includes for every premise, claim or Non-argument text:

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

and for every argument relation:

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

Moreover, we provide an additional JSON file including the following labels, that were not marked in the original transcript:

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

These latter annotations could be useful particularly for a financial application scenario. For instance, analyst-based prediction systems, Q&A sentiment correlations and others.

An example of Label Studio interface, along with the corresponding JSON and .ann files, are presented in Appendix d. Note that the part of argument quality will be discussed in the next chapter (Chapter 5).

### 4.2.4 *The FinArg Corpus Statistics*

Table 4.1 shows statistics about our annotated data distributions. The number of documents represents the number of different analysts, as we have clarified earlier. However, an analyst usually has the right of two different questions. They may be formulated together [9], or stated with a follow-up question. Moreover, for some questions, two of the company representatives may answer, individually. Therefore, the number of annotated answers can be (and it is) more than the number of questions. In particular, this case is observed in FB and MSFT data (see Table 4.2).

The found proportion between claims and premises is also common in argumentation and confirms the findings of [MM11; SG14a] that claims are usually supported by several premises for ensuring a complete and stable standpoint. Additionally, the proportion between support and attack relations is intuitive, since discussing the opposite standpoint, as a preemptive self-defense, is less commonly used in argumentation comparing to the direct supporting premises. There are 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. Expressly, we want to avoid counting them as new arguments. Thus, we calculate the percentage of "unlinked" relatively to the total number of premises and claims. Furthermore, Table 4.2 shows a detailed version of the classes distributions per different companies.

9 This will be counted as one by the statistics script.

Table 4.1: Corpus statistics and class distribution

| Type | Count | % |
|---|---|---|
| **Earnings calls** | 80 | - |
| **Documents** | 839 | - |
| **Questions** | 1621 | - |
| **Answers** | 1859 | - |
| **Premises** | 5098 | 35.903% |
| **Claims** | 4639 | 32.671% |
| **Non-argument** | 4462 | 31.424% |
| **Support** | 4843 | 98.394% |
| **Attack** | 79 | 1.605% |
| **Unlinked** | 1786 | 18.342% |

Table 4.2: Distribution per company where FB: Facebook, AAPL: Apple, AMZN: Amazon, MSFT: Microsoft

| Type | FB | AAPL | AMZN | MSFT |
|---|---|---|---|---|
| **Earnings calls** | 20 | 20 | 20 | 20 |
| **Documents** | 264 | 138 | 239 | 198 |
| **Questions** | 421 | 431 | 374 | 395 |
| **Answers** | 489 | 431 | 374 | 565 |
| **Premises** | 1718 (38.42%) | 1006 (31.03%) | 1148 (34.64%) | 1226 (38.63%) |
| **Claims** | 1427 (31.91%) | 1078 (33.26%) | 1077 (32.49%) | 1057 (33.31%) |
| **Non-argument** | 1326 (29.65%) | 1157 (35.69%) | 1089 (32.86%) | 890 (28.04%) |
| **Support** | 1645 (98.73%) | 926 (96.35%) | 1073 (99.35%) | 1199 (98.68%) |
| **Attack** | 21 (1.26%) | 35 (3.64%) | 7 (0.64%) | 16 (1.31%) |
| **Unlinked** | 375 (11.92%) | 481 (23.08%) | 475 (21.34%) | 455 (19.92%) |

## 4.3 PARSING THE ARGUMENTATIVE TEXT AND ARGUMENT COMPONENTS

Our novel corpus *FinArg* lays the foundation of the automatic parsing of argumentative text and argument components in financial earnings calls. In this section, we report the performance of our argument mining model presented in Chapter 3 on both AM tasks.

In terms of argument identification task, Table 4.3 shows that we got an accuracy of 0.85 and F1-score of 0.81, which are comparable to this model outcomes on *Student essays* [SG14a] and *User-generated web discourse* [HG17] presented in [Alh+21b], as shown in Tables 3.2 and 3.3. Similar conclusions are reached on the argument unit classification task (see Table 4.4).

Moreover, we consider those experiment as a strong baseline for future work, where different points can be improved. For instance, the unit segmenta-

tion process and the feature engineering part for SVM classifier to be more indicative of this text properties.

Table 4.3: Evaluation of argument identification task on the FinArg dataset

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM | 0.8234 | 0.8124 | 0.7437 | 0.7650 |
| DistilBERT | 0.8434 | 0.8459 | **0.7899** | 0.8102 |
| Stacking | **0.8570** | **0.8594** | 0.7872 | **0.8114** |

Table 4.4: Evaluation of argument unit classification task on the FinArg dataset

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM | 0.6800 | 0.6795 | 0.6779 | 0.6782 |
| DistilBERT | 0.7514 | **0.7517** | **0.7523** | 0.7513 |
| Stacking | **0.7529** | **0.7517** | **0.7523** | **0.7517** |

To sum up, our preliminary 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 amount to the granularity of data needed for future work on the prediction of analysts' post-call recommendations.

## 4.4 CONCLUSION

In this chapter, we contribute to the (1) theory, (2) data and (3) evaluation aspects of argumentation structure in the financial domain by: (1) proposing a micro-structure argumentation scheme for modelling arguments presented in company representatives' responses during the earnings conference calls, (2) working 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 the size of 839 documents, and (3) evaluating this reliability of this data by measuring different inter-annotator agreement as well as examining our stacking approach as an automatic parser.

We conclude that this data has many potentials to foster the research in computational argumentation since it is the first dataset that considers this type of text, and it covers all of non-argumentative, argument units, and relations. Moreover, it composes the core-stone of a new research foundation for financial document processing in the FinTech interdisciplinary field. A wide-range of applications can be developed and further invested, including but not limited to:

- Volatility prediction, which is a key to the risk perspective.
- Help investors to make more informed decisions by efficiently marking the argumentative parts in the discussions of ECCs.

- Understanding correlations between the executives' given arguments and analysts' recommendations.

We further investigate on the argument quality aspect in Chapter 5.

# 5

# ARGUMENT QUALITY ASSESSMENT IN EARNINGS CONFERENCE CALLS

After dealing with the argument structure task in Chapter 4. We now aim at answering the third research question of this thesis:
***RQ3***: 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?

Most of preceding work in computational argument quality (CAQ) focuses on assessing the overall quality or only a specific concept of AQ [Lau+20].

However, not surprisingly, Eemeren et al. [EH02] linked the speaker's strategy of maneuvering to the 'audience demand'. Hence, we have to consider the market analyst and investors expectations while listening to the earnings calls. That is to say, while rhetorical figures may play a more considerable rule in legal text [Sau94], real-world values speak for themselves in a financial era. Therefore, it is crucial to define the argument quality dimensions with respect to the features of this genre of text as well as the market point of view.

In this chapter, we tackle this research gap by conducting a comprehensive study on earnings calls and CAQ state of the art. Investigating on the same *FinArg* corpus, we have introduced in Chapter 4, the contributions of this chapter are:

- **Theory:** Based on CAQ literature and the financial perspective, we propose different quality dimensions, considering both the type of argumentation and the overall argument attributes.
- **Data**: We conduct the related annotation study and produce *FinArgQuality*: the first financial corpus annotated with AQ scores.
- **Evaluation:** We further propose a machine learning approach to the automatic assessment of one of our AQ dimensions: argument strength.

Our proposed assessment methodology, dataset, and evaluation approach can serve as strong baselines for future work.

## 5.1 RELATED WORK

Our work is closely related with two existing lines of research:

### 5.1.1 *Argument Quality Assessment in Computational Argumentation*

Delving into the rich realm of argumentation theories, various quality proposals have been introduced. The persuasion itself is a product of different factors, as we have seen in Section 2.4. Nevertheless, researchers in computational argumentation looked for practical, yet considerable definitions of argument quality. They further faced this problem with different methodologies of assessment. Figure 5.1 displays an overview of the literature approaches for evaluating argument quality:

- **Rating:** Point-wise versus Pair-wise approach. Meaning that, either an absolute rating of the argument (e.g., ranking the strength of a student essay [PN15]), or a relative rating of it in comparison with another argument (e.g., which argument is more convincing by [HG16]).
- **Level of Granularity:** With respect to the level of granularity, we can distinguish methods that estimate the quality of the complete argument (e.g., [FSB15]) versus, the quality of its particular components (e.g., [Rin+15]). Furthermore, some scholars explored the interactions into debate context. For instance, [Tan+16] studied the persuasion on the Reddit platform as a function of interaction dynamics between the opinion holder and the counterargument provider. They tried, hence, to define the wining argument using the interaction patterns.
- **Method of Assessment:** The literature reported mostly direct classification (regression) models (e.g., [SG17b; Lau+20]), with some indirect attempts. For instance, [WW20] investigated on a set of linguistic features that reflect the argument quality instead of considering the original text. Similarly, Gurcke et al. [GAW21] aimed at assessing the sufficiency of arguments through conclusion generation. However, as expected, direct methods outperform their peers.

This discussion should give a bird's-eye view on the diversity of computational argument quality field. Nevertheless, we have to point out that there is still no consensus whether argument quality should be assessed from a normative theoretical or practical descriptive view [All16] (see Section 2.4).

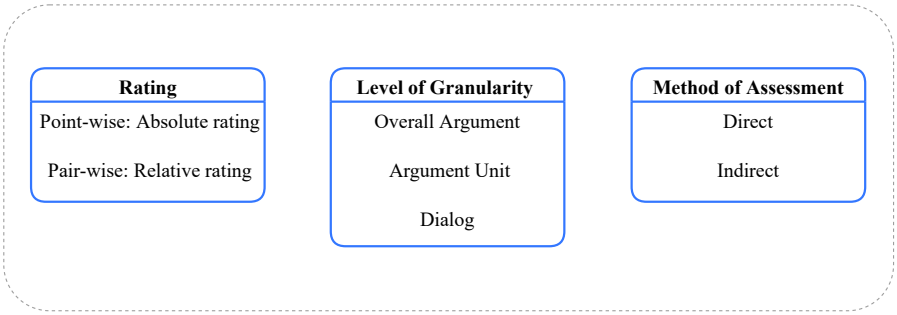



Figure 5.1: A taxonomy of state-of-the-art computational argument quality assessment.

In their interesting book [VEGE04], van Eemeren and Grootendorst, supported the need of a "normative pragmatics" theory of argumentation. Moreover, [Wac+17a] proved that even those "spontaneously" defined quality metrics, are still "represented by theory". This means that humans inherently can comprehend and judge arguments correctly, even without having the full knowledge or formality of experts. Therefore, they suggested simplifying theory to increase its applicability to the real-world arguments.

Based on all of that, we aim in our study to define a set of metrics that highlight the quality of company executives' arguments in a pragmatic, yet quantifiable manner.

Furthermore, while most studies treat the argument in a holistic manner, Walton [Wal96] 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 argument's premises. We also extend that to all argumentative units, so we consider the varieties of argument claim as well. We provide further discussions all across our quality dimensions.

### 5.1.2 *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 [CHC21a; ZRH20] evaluated the forecasting skills of investors, [QY19] analyzed the managers' speech with the goal of predicting the financial risk. Consequently, various data-source have been studied, including social media [Alh+21c], earnings conference calls [KS19], public news articles [Aga20], and others.

We present in the following some related work that is directly linked to our proposed quality metrics:

Zong et al. [ZRH20] used the Linguistic Inquiry and Word Count (LIWC) lexicon [TP10] 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, [RRP19] 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" [ZRH20] and "hedging" [KS19] are only indicators of "the lack of commitment to the content of the speech" [PH14]. 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.

## 5.2 THE PROPOSED DIMENSIONS OF ARGUMENT QUALITY

Given that the criteria of what is a good argument depends on the goal orientation [WW20; JB06], 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[1]. The rating follows the point-wise approach, and looks at each argument from two levels:

### 5.2.1 *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 complete argument. In other words, considering the argument claim and premises as well as the relations between them.

- **Strong**
  Persing et al. [PN15] labeled the strength of a student essay (and not of one argument) using a scale 1.0 to 4.0 with 0.5 increments. On the other hand, [Car+18] 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 5.1 represents the rubrics for rating the argument strength.

Table 5.1: Strength dimension of an argument

| Score | Description |
|---|---|
| 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. |

1 https://www.wiwi.uni-passau.de/en/financial-data-analytics

- **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 5.2 displays our hints to label persuasiveness across arguments.

Table 5.2: Persuasiveness dimension of an argument

| Score | Description |
|---|---|
| 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. |

- **Specific**
  Carlile et al. [Car+18] 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. [PH14] 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 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 5.3.

Table 5.3: Specific dimension of an argument

| Score | Description |
|---|---|
| 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. |

- **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? (see Table 5.4).

Table 5.4: Objectivity dimension of an argument

| Score | Description |
|---|---|
| Objectivit-1 | A logical argument supported by verifiable evidences. |
| Objective-0 | A subjective or biased argument based on particular views and 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 [CHC21a; Che+18; UzZ+13]. On the other hand, in a business communication study, Crawford et al. [CC18] 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 [ZRH20] 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 5.5.

Table 5.5: Temporal-history dimension of an argument

| Score | Description |
|---|---|
| 3 | Recent: during this quarter. |
| 2 | Short-past: up to 2 quarters. |
| 1 | Mid-past: half to one year. |
| 0 | Long-past: more than 1 year. |
| -1 | Not mentioned: if there is no explicit time indicator. |

### 5.2.2 *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 [BMB10]. 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 Figure 5.2.

#### 5.2.2.1 *Types of Claims*

Clarile et al. [Car+18] 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 [CC18].

  > 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.

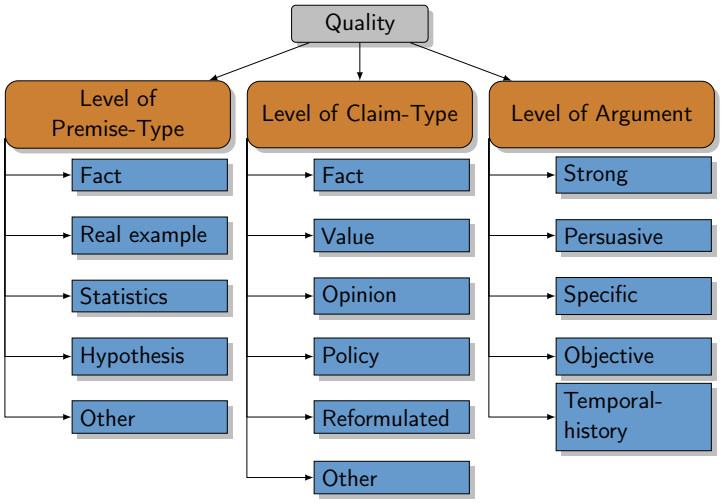



Figure 5.2: Our quality dimensions at the levels of argument and argument units

> 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 [VSD12], 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.

#### 5.2.2.2 *Types of Premises*

Similarly to claims forms, and motivated by the works of [AK+16a; Car+18], we set the following premise types:

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

> 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 [AK19].

  > 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 [Che+18].

- **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 give us 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 [WRM08] (cf. Section 2.2.1).

## 5.3 FINARGQUALITY CORPUS CREATION

### 5.3.1 *Annotation Study*

The annotation of the argument quality metrics (*FinArgQuality* corpus) was done simultaneously with the argument components and relations (*FinArg* corpus, cf. Section 4.2).

Hence, this corpus covers the earnings calls of four companies (Amazon, Facebook[2], Microsoft, and Apple) during 2015-2019. Moreover, the annotators do not have pre-defined arguments to rate. Instead, they have to label each of argument units, relations, and quality scores.

Each argumentative unit is represented in a sub-sentence (clause) at minimum, and one sentence at maximum. Each unit refers to one type of premise or claim, and no intersection is allowed. The overall argument quality criteria are judged while considering all of its premises, claim, and relations.

Appendix c describes the detailed annotation guidelines of argument quality assessment. Furthermore, Appendix d shows an example of the annotation process, using Label Studio interface, conjointly with representative segments of the corresponding JSON and .ann files.

Similarly to *FinArg*, after the annotations are complete, we proceed through a data cleansing phase, where erroneous data points are detected and reported to the annotators for correction. These steps are repeated until all issues are resolved.

### 5.3.2 *The FinArgQuality Inter-annotator Agreement*

As we have aforesaid, the annotation of this corpus was held simultaneously with *FinArg*. While this decision save us time, it still leads to lower agreement on the argument quality, since the argument components could be different in the first place (inheritance bias).

We calculate three evaluation measures: Cohen's Kappa [Coh60], Krippendorff's Alpha [Kri18], and Fleiss Kappa [FLP13]. However, all lead to similar results, given that we have only two of our annotators for each document from the validation set. Therefore, we report only Cohen's kappa results in this thesis. For all data, we measure the agreement separately for each pair of annotators, and report the average. Table 5.6 shows that we obtained fair to substantial agreements [LK77].

However, the results show some unevenness in the quality of annotators, and their agreements on particular documents. One reason could be the obvious differences between the managers' attitude of speech. For example, those who tend to use long sentences, make it harder to get an agreement between the two annotators on the unit boundaries or label. Nevertheless, despite the different number of documents per each company (cf. Table 4.2), the data distribution across them does not reflect any outliers, as we will see in Section 5.3.3.

Moreover, similarly to argument components and relations (cf. Section 4.2.2.2), the main source of disagreement is the missing of unit boundaries, and the multiple possible interpretations of argument structure [WR03; SG14a; Hen00]. This, definitely, applies to rating argument quality, which is even more inherently subjective [Wac+17a]. 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

2 Recently Meta.

degrees of specificity, strength, and persuasiveness. Thus, extending guidelines with those cases would improve further annotations.

Table 5.6: 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 |

In a similar study, Wachsmuth et al. [Wac+17b], introduced the *Dagstuhl15512 ArgQuality Corpus* for studying argumentation quality based on their developed taxonomy of 15 dimensions (cf. Figure 2.7) on a 3-point scale (low, medium, high). They reported Krippendorf's $\alpha$ of all annotators ranging from 0.174 to 0.447 only.

### 5.3.3 *The FinArgQuality Corpus Statistics*

In this section, we present the detailed statistics and distributions of *FinArgQuality* corpus. Table 5.7 composes an overview on the data size and ratio of argumentative text, both in terms of total, average per document [3], and average per company. A document has, in average, 12 in-argument (argumentative) sentences and 6 out-of-argument sentences. In other words, 68% of the dataset consists of argumentative components. This confirms the value of studying this kind of data by the means of argument mining.

The overall quality dimensions are described in Table 5.8 in total, and per company. The percentages are based on the total number of arguments. We can observe similar distributions over the different companies, which indicates that our dataset can be a real representation of the population.

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 of arguments, and the persuasion strategies managers often use during the earnings calls, as highlighted by Crawford [CC18].

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.

3 As a reminder, the document represents a single analyst's questions and their answers in one earning call

Table 5.7: Proportions of argumentative and non-argumentative units over *FinArgQuality*. The average is presented along with its standard deviation

| | **Attribute** | **Count** | [%] | **Avg. per doc** | | | **Avg. per company** | | |
|---|---|---|---|---|---|---|---|---|---|
| Sentences | In-argument Sentences | 9693 | 68.53 | 12 | ± | 6 | 2423 | ± | 423 |
| | Out-of-argument Sentences | 4453 | 31.47 | 6 | ± | 4 | 1113 | ± | 158 |
| Tokens | In-argument Tokens | 244253 | 78.84 | 297 | ± | 155 | 61063 | ± | 13437 |
| | Out-of-argument Tokens | 65537 | 21.16 | 82 | ± | 78 | 16384 | ± | 4796 |

On the other hand, Table 5.9 presents the detailed statistics with respect to the claim and premise types. 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 data collected, 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. In a similar analysis, Villalba and Saint-Dizier [VSD12] show how "a number of evaluative expressions with a 'heavy' semantic load receive an argumentative interpretation".

Table 5.8: Statistics of *FinArgQuality* dimensions

| Company / Quality dimension | 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 |
| 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 |
| 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 |
| 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 |
| TEMPORAL-HISTORY -1 | 338.0 | 15.48 | 288.0 | 13.19 | 733.0 | 33.56 | 408.0 | 18.68 | 1767.0 | 80.91 |
| TEMPORAL-HISTORY 0 | 26.0 | 1.19 | 18.0 | 0.82 | 12.0 | 0.55 | 4.0 | 0.18 | 60.0 | 2.75 |
| TEMPORAL-HISTORY 1 | 54.0 | 2.47 | 43.0 | 1.97 | 7.0 | 0.32 | 10.0 | 0.46 | 114.0 | 5.22 |
| TEMPORAL-HISTORY 2 | 24.0 | 1.10 | 41.0 | 1.88 | 17.0 | 0.78 | 11.0 | 0.50 | 93.0 | 4.26 |
| TEMPORAL-HISTORY 3 | 48.0 | 2.20 | 45.0 | 2.06 | 40.0 | 1.83 | 17.0 | 0.78 | 150.0 | 6.87 |

## 5.4 ANALYSIS OF CORRELATIONS BETWEEN QUALITY DIMENSIONS

The analysis of potential correlations between argument quality dimensions, at both argument and argument unit level, is useful from different perspectives.

Table 5.9: Statistics of Claims and Premises types. The average is presented along with its standard deviation

| | Attribute | Count | [%] | Avg. per doc | Avg. per company |
|---|---|---|---|---|---|
| Claims | CLAIM-Fact | 2001 | 43.38 | 3 ± 2 | 500 ± 93 |
| | CLAIM-Opinion | 1672 | 36.25 | 2 ± 2 | 418 ± 64 |
| | CLAIM-Reformulated | 850 | 18.43 | 2 ± 1 | 212 ± 56 |
| | CLAIM-Policy | 45 | 0.99 | 1 ± 0 | 11 ± 5 |
| | CLAIM-Value | 28 | 0.60 | 1 ± 0 | 7 ± 3 |
| | CLAIM-Other | 17 | 0.37 | 1 ± 0 | 4 ± 3 |
| Premises | PREMISE-Fact | 3624 | 71.37 | 5 ± 3 | 906 ± 303 |
| | PREMISE-Statistic | 691 | 13.60 | 2 ± 1 | 173 ± 92 |
| | PREMISE-RealExample | 496 | 9.77 | 2 ± 1 | 124 ± 53 |
| | PREMISE-Hypothesis | 46 | 0.91 | 1 ± 0 | 12 ± 5 |
| | PREMISE-Other | 221 | 4.35 | 2 ± 1 | 55 ± 24 |

First, it provides us with a clear idea about the uniqueness of each of our quality dimensions. Second, correlations can afford indicators and hints for the machine learning model design.

Figure 5.3 reflects the linear Pearson's correlations. Apparently, the *Strong* attribute is the most useful in understanding the persuasiveness of an argument, since they are the most correlated.

The Strength itself is correlated with number of supporting premises (p=0.5) which corresponds to our annotation guidelines.

Specificity is also correlated with argument strength and persuasiveness. However, that does not imply causality. For example, an argument can be specific to a particular question (*Specific*: 2), but the premises cited to support the claim are weak (*Strong*: 0).

In general, different premise-types are positively correlated with the overall quality, whereas, claim-type is only considered for *Objectivity* judgments. This suggests using the premise type and relation type for the tasks of predicting overall quality, which we will further examine in the following Section 5.5.2.

In addition, the low correlation between different dimensions emphasizes their unique role in the overall quality assessment. Thus, we have to consider all of them to produce a fair estimation of the quality.

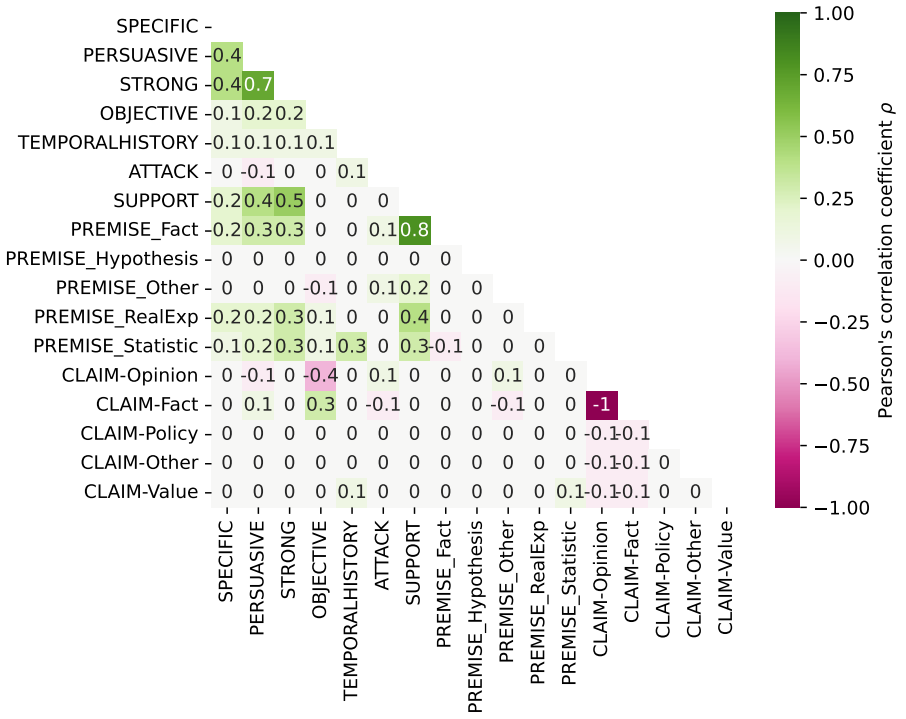



Figure 5.3: Pearson's correlation between argument quality dimensions

## 5.5 COMPUTATIONAL ARGUMENT QUALITY ON FINARGQUALITY - ARGUMENT STRENGTH

The strength of an argument is a key dimension to represent its logical quality [Wac+17a; PN15]. A reliable computational assessment of argument strength enables systems to identify well-supported arguments. Moreover, in our data, it is the most associated dimension with argument's persuasiveness (cf. Section 5.6). Therefore, we propose, in this thesis, a deep learning model for predicting the strength score of arguments. Argument strength is measured using ordinal scores ranging from 0 to 2. Hence, we have a multi-class classification task to be solved.

### 5.5.1 *Data Pre-processing*

Towards our goal of automatic argument quality assessment, we transform the multiple text annotation documents into a processed CSV file where each row represents one data point (one argument). For each of them, we include the claim, linked premises, relation types, claim types, and premises types, as well as quality metrics annotations. For any potential traceability need, we also include metadata, such as company name, year, quarter, annotator ID, and the original file ID.

#### REMOVING IRRELEVANT DATA POINTS

The strength score is rated at the argument level. In light of this, only

full arguments with a complete annotation are taken into account. Namely, out-of-argument sentences, unlinked claims as well as unlinked premises are neglected. This introduced a downsizing effect, as we can see in Tables 5.10 and 5.11 for claim and premise types, respectively. Indeed, 96% of premises and 47% of claims are linked (18.43% of unlinked claims are reformulated). Nevertheless, there are almost no changes in the distribution of claim types, premise types, or relation types. Overall, we have 4899 premises associated with 2184 claims, which results in an average of 2.24 ± 1.68 premises per claim (argument). Thus, we still have an imbalanced data issue.

Table 5.10: Statistics of claim labels in *FinArgQuality* after pre-processing

| Company / Claim label | FB | | AMZN | | MSFT | | AAPL | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Count | [%] | Count | [%] | Count | [%] | Count | [%] | Count | [%] |
| CLAIM-Fact | 319.0 | 14.61 | 195.0 | 8.93 | 473.0 | 21.66 | 248.0 | 11.36 | 1235.0 | 56.55 |
| CLAIM-Opinion | 157.0 | 7.19 | 220.0 | 10.07 | 322.0 | 14.74 | 196.0 | 8.97 | 895.0 | 40.98 |
| CLAIM-Reformulated | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| CLAIM-Policy | 10.0 | 0.46 | 5.0 | 0.23 | 6.0 | 0.27 | 2.0 | 0.09 | 23.0 | 1.05 |
| CLAIM-Value | 2.0 | 0.09 | 8.0 | 0.37 | 1.0 | 0.05 | 1.0 | 0.05 | 12.0 | 0.55 |
| CLAIM-Other | 2.0 | 0.09 | 7.0 | 0.32 | 7.0 | 0.32 | 3.0 | 0.14 | 19.0 | 0.87 |

Table 5.11: Statistics of premise labels in *FinArgQuality* after pre-processing

| Company / Premise label | FB | | AMZN | | MSFT | | AAPL | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Count | [%] | Count | [%] | Count | [%] | Count | [%] | Count | [%] |
| PREMISE_Fact | 698.0 | 14.25 | 577.0 | 11.78 | 1365.0 | 27.86 | 888.0 | 18.13 | 3528.0 | 72.01 |
| PREMISE_Statistic | 270.0 | 5.51 | 180.0 | 3.67 | 140.0 | 2.86 | 46.0 | 0.94 | 636.0 | 12.98 |
| PREMISE_RealExample | 54.0 | 1.10 | 100.0 | 2.04 | 129.0 | 2.63 | 194.0 | 3.96 | 477.0 | 9.74 |
| PREMISE_Hypothesis | 3.0 | 0.06 | 16.0 | 0.33 | 13.0 | 0.27 | 12.0 | 0.24 | 44.0 | 0.90 |
| PREMISE_Other | 46.0 | 0.94 | 81.0 | 1.65 | 19.0 | 0.39 | 68.0 | 1.39 | 214.0 | 4.37 |

#### DATA IMBALANCE

The review of data statistics in Table 5.8 reveals an "expected" imbalance issue. The majority class corresponds to Strong-1 with a percentage of 65%. However, this is common in AQ datasets.

For instance, Stab et al. [SG17b] reported 681 (66.2%) sufficient to 348 (33.8%) insufficient arguments in their student essays corpus.
Likewise, in the corpus of Persing and Ng [PN15] annotated with strength scores [4], among the 1000 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.

### 5.5.2 *Method*

BERT [Dev+18] has set a new state-of-the-art performance on various *sentence-classification* tasks. Despite the release of different language

4 As aforementioned, using a scale 1.0 to 4.0 with 0.5 increments, giving a total of seven values

representation models, BERT is still powerful in automating language *understanding*. This may be due to the fact that BERT itself has 24 layers (transformer blocks) and 345 million parameters. It is a large and complex model with a bidirectional functionality that empowers conceptual understanding and the detection of long term dependency.

Yet, we have tested XLNet [Yan+19b], a generalized autoregressive model for language understanding, which has a similar architecture to BERT. They both achieved the same F1-score (48%), but the execution time of XLNet was three times more than BERT. No significant improvement was achieved to cover this execution cost. Similarly, in an extensive study by Facebook AI [Liu+19], they proved that BERT can match or exceed the performance of every model released after it, including XLNet. Moreover, the access to some more recent language models like the Generative Pre-trained Transformer 3 (GPT-3) is only possible through an API (Application Programming Interface), which is not relevant for a research setting. Therefore, we decide to build further on BERT.

Moreover, BERT is in line with previous works [Tol+19; Gre+20], which can provide us with a baseline for comparison.

Nevertheless, in our experiments, our focus is not about running the best model. Rather, we tackle our argument classification task through its characteristics and requirements.

In summary, considering the argument strong attribute, (1) the data is unbalanced, (2) the annotation is at the whole argument level (premises and claim have to be considered), (3) and the types of argument units and relations could be useful in the learning process. We address those challenges in the following one by one.

#### 5.5.2.1 *Baseline: Bert*

According to the annotation of the FinArgQuality dataset, the Strong dimension is one of the quality dimensions defined at the argument level (cf. Table 5.1). Consequently, we need to assess an entire argument (the claim along with its linked premise(s)).

Transformers often have limits on input length, such as Bert's maximum of 512 tokens [Din+20]. In our case, the input does not need to be truncated as our arguments, on average, contain 81 tokens.

To input the complete argument, we separate the different argument components using the Bert predefined special token (**[SEP]**). This concatenation approach is widely adopted in NLP tasks where some additional information is needed. For example, to chain the argument with the corresponding topic, as in [Rei+19], and similarly to concatenate the claim with its context, as in [DLC20]. Hence, we have the following:
**Input** = *claim* [SEP] $premise_1$ [SEP] $premise_2$ [SEP]... $premise_n$
**Output** = 0,1 or 2

#### 5.5.2.2 *Improvement 1: Bert with special separator token*

Using the Bert delimiter [SEP] unifies the role of argumentative units. Meaning that, the model does not recognize whether the separated unit is a premise or a claim. However, we can easily define a new delimiter by adding it to the special tokens list of Bert. In this context, Lopez et al. [Lop+20] proposed an approach to mark the answer-start ([ANSS]) and the answer-end ([ANSE]) for a question generation task. This approach leads to outstanding results and outperformed complex Seq2Seq methods. Likewise, and for each argument unit, we establish two tokens to mark its starting and ending points. We, hence, test if that would contribute positively to the learning process, as follows:
**Input** = [cl_text] *claim* [/cl_text] [pr_text] $premise_1$ [/pr_text] ... [pr_text] $premise_n$ [/ pr_text]

#### 5.5.2.3 *Improvement 2: Bert with class weights*

To overcome the imbalanced data issue (cf. Table 5.8), we re-weight the loss using inverse class frequency.

Since the data is highly imbalanced, and we want to treat all classes equally, we use the macro-F1 score for evaluation. This is also suggested by Stab et al. [SG17a].

Therefore, to determine the optimal weights, we conduct a statistical analysis of the macro F1-score while varying the class weight of the minority classes (0 and 2).

#### 5.5.2.4 *Improvement 3: Bert with categorical features as text*

Estimating the quality of an argument is not a simple classification problem, since it inherently depends on the context and understanding of the argument. Fortunately, we have in our *FinArgQuality* data, some additional information that can be straightforward indicators of the argument strength, which we consider as *categorical features*: premise type, claim type, and relation type.

We incorporate all three features together at this step. In the first instance, we include those features as text. In other words, the string categorical features (premise type, claim type, and relation type) are concatenated to the argument and passed as input to Bert. Similar to improvement step 2, we define new separator tokens to consider the types. The following symbolizes an example:
**Input** = [cl_text] It's a very rapidly expanding country. [/cl_text]
[cl_type] Claim-Opinion [/cl_type]
[r_type] Support [/r_type]
[pr_text] Constant currency growth was 48%.[/ pr_text]
[pr_type] Premise-Statistic [/pr_text]

Moreover, we conduct an *exhaustive feature selection* method. This could be highly computationally expensive, since it trains on all possible sets of features

[Nag+15]. However, given that we have only three categorical features, (total of eight combinations), this should not pose a problem for us.

#### 5.5.2.5 *Improvement 4: Bert with encoded categorical features*

Incorporating categorical features (premise/claim/relation types) improved the model performance. Yet, involving them as a text in the input string has a disadvantage on the final generated embeddings. That is because having them as part of the argument adds incorrect context to it [LKB20]. Moreover, those values (e.g., premise-Statistic) should have a static impact on the model knowledge, and should not be embedded differently based on the argument text itself.

To tackle this problem, we suggest to:

1. Convert the categorical features separately to an encoded numerical vector.
2. Concatenate that to the contextual embedding of the argument text (generated using Bert tokenizer).
3. Insert the output chained vector to the Bert classification layer.

The conversion of categorical features to numerical inputs can be accomplished using several techniques. Among them, we have Ordinal Encoding and One-Hot Encoding. Since the types of claims, premises, and relations, do not imply any order relation, we apply One-Hot Encoding to avoid any deceptive information for the model.

On the other side, the argument is inserted to the Bert base, along with the special separators we defined earlier. Consequently, each token in the input sequence is represented by a hidden state vector. A variety of NLP tasks such as question answering [Alh+22a; Yan+19a], sequence classification [Sun+19], and sentiment analysis [Xu+19] are performed using these hidden states from Bert's last layer. However, in classification tasks and as explained in [Dev+18], only the [CLS] hidden state representation is used as input to the classification layer.

Hence, in the final stage, we assemble the [CLS] hidden state with the encoded vector of categorical features and feed them into the final classifier.

The overall architecture is illustrated in Figure 5.4. All in all, we use Bert architecture, we add the encoded extra categorical features to the [CLS] hidden state output of Bert. On top of both, a classifier layer takes the combined vector as input and outputs a vector of size 3 (number of labels).

### 5.5.3 *Evaluation*

Throughout this section, we discuss the experimental setup and results of our argument's strength classification model on the FinArgQuality dataset.

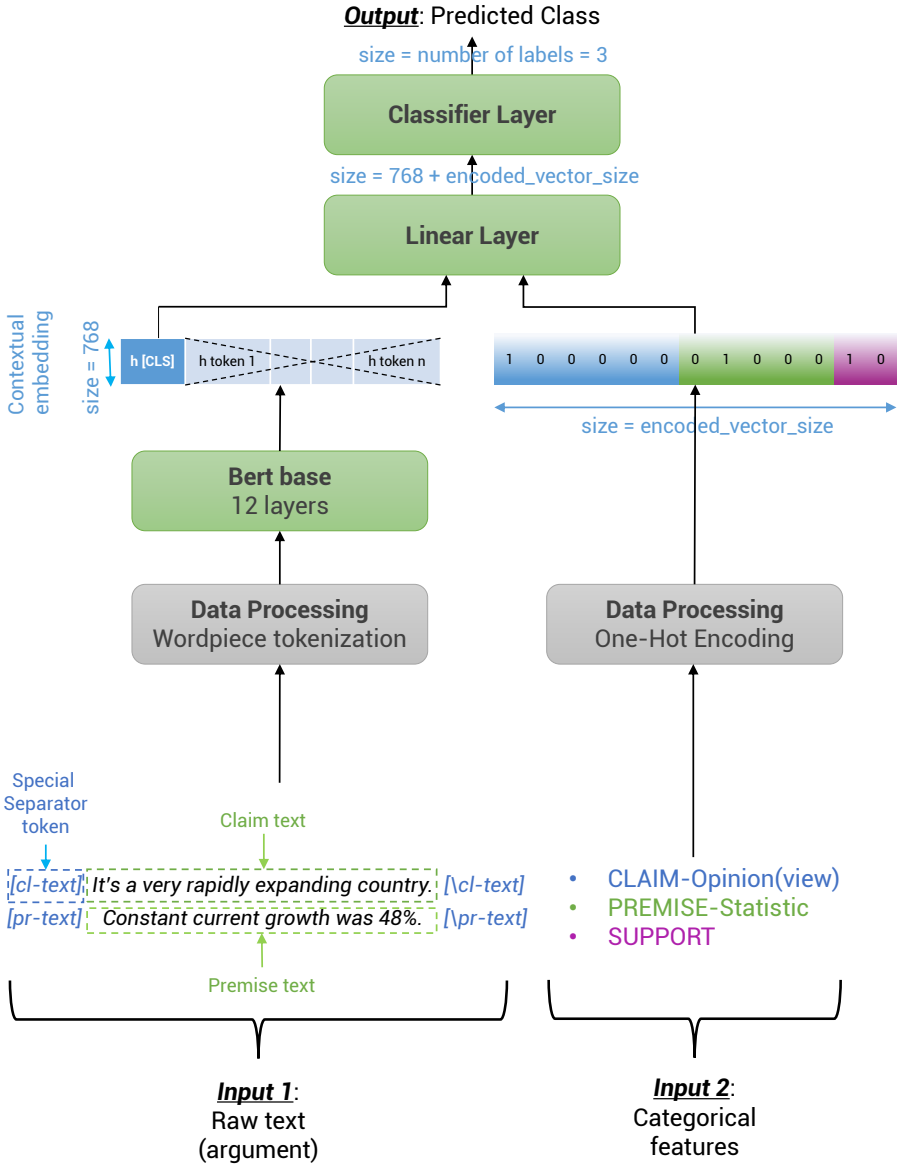



Figure 5.4: Model architecture for Bert with encoded categorical input features.

EXPERIMENTAL SETUP

Considering the limited size of the data, we conduct a model selection experiment using iterative stratified sampling over 10-fold cross-validation [STV11; SK17]. The iterative stratified sampling method is used to obtain representative folds of the data while considering not only the distribution of the target class (Strong) but also the distribution of other classes.

In our particular case, we make the split while having approximately an equal distribution of the strong class, companies, source documents, premise types, and claim types. Apparently, the distribution among the folds could not be 100% precise, since it is not simple to satisfy all constraints. The more classes we include in the split process, the more constraints we have.

For all experiments, we fix the batch size to 8, the learning rate to 5 e-5, and the number of epochs to 3. In addition, even though the data is imbalanced, we consider all classes are equally important for us. Thus, we adopted macro-averaging in our experiments [LLS09].

Furthermore, to test the significance of our model's results, we use Student t-test [NB07] with p = .05.

RESULTS

Table 5.12 summarizes the results of all of our experiments. The best outcome was obtained using Bert with encoded categorical features. We present, in the following, a detailed discussion of each improvement step.

#### 5.5.3.1 *Baseline*

By fine-tuning the original Bert model on our data, while using the [SEP] to concatenate argument components, we got 74% of accuracy, with a 48% F1-score.

#### 5.5.3.2 *Bert with special separator token*

In this step, we define the special tokens [cl-text] and [pr-text] to indicate the start of claim and premise texts, respectively. This leads to an enhancement of 3% w.r.t. the accuracy and 2% w.r.t. the F1-score.

According to the Student t-test, the results ($p = 0.02 \leq 0.05$) demonstrate a statistically significant improvement.

Table 5.12: Evaluation of the different examined models, on *FinArgQuality*, where Sem stands for standard error of the mean.

| | **Macro-Precision** | | | **Macro-Recall** | | | **Macro-F1 score** | | | **Accuracy** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | ± | Sem | Mean | ± | Sem | Mean | ± | Sem | Mean | ± | Sem |
| Bert (Baseline) | 0.48 | ± | 0.02 | 0.48 | ± | 0.01 | 0.48 | ± | 0.01 | 0.74 | ± | 0.01 |
| Bert, special separator token | 0.50 | ± | 0.01 | 0.51 | ± | 0.00 | 0.50 | ± | 0.00 | **0.77** | ± | 0.01 |
| Bert, class weight | 0.55 | ± | 0.02 | 0.59 | ± | 0.02 | 0.55 | ± | 0.02 | 0.67 | ± | 0.02 |
| Bert, features as text | 0.56 | ± | 0.01 | 0.60 | ± | 0.02 | 0.56 | ± | 0.01 | 0.71 | ± | 0.01 |
| Bert, One-Hot Encoding | **0.61** | ± | 0.02 | **0.63** | ± | 0.02 | **0.61** | ± | **0.02** | 0.74 | ± | 0.01 |

However, we suggest that the model must be fine-tuned on a sufficient amount of data in order to learn the meaning of each new defined token. When training data is insufficient, it is recommended to use the predefined special token [SEP]. Henceforth, and in all the upcoming experiments, we use our delimiter tokens.

#### 5.5.3.3 *Bert with class weights*

According to Table 5.8, and for the strength dimension, 65% of the arguments are categorized under label 1. The size of class 1 is approximately 8 times the size of class 0 and twice the size of class 2. To tackle this imbalance issue, we adopt the class weight technique. We have tested, using 10-folds cross validation, a total of 12 combinations of class weights. We vary the weight of class 0 from 5 to 11 and of class 2 from 1 to 2. Whereas, class 1 is the majority class, so its weight is set to 1, as shown in Figure 5.5.

The weights set [7,1,2] (corresponding to classes [0,1,2]) returns the best model performance. Applying these weights, all of the precision, recall and F1-score are improved comparing to the previous step 5.5.3.2. However, increasing the weight of minority classes ensues decreasing the weight of the majority class. Consequently, the classes with more representation are classified less accurately. Thus, the overall accuracy decreases.

Nevertheless, since the outcomes are close to each other. We calculate the Student t-test, while compared to Bert with special tokens. It yields ($p = 0.15 > 0.05$) for the weights [7,1,2] and ($p = 0.01 < 0.05$) for [8,1,2]. Accordingly, we select weight [8,1,2], which results in the second-best score but with a significant improvement.

Henceforward, all experiments are conducted using new special tokens and with the weights [8,1,2].

#### 5.5.3.4 *Bert with categorical features as text/One-Hot encoding*

As we have described in Sections 5.5.2.4 and 5.5.2.5, we aim at integrating the categorical features (claim/premise/relation types) in the learning process.

The results of all combinations of categorical features included as text and as a One-Hot encoded vector are represented in Table 5.13.

For features as text, the outcome F1-score ranges from 49% to 56%. The highest model performance (56%) is obtained when including only the premise type. For all other sets of features, the macro-F1 score is equal to or less than the macro-F1 score of the model without features (55%). For instance, we

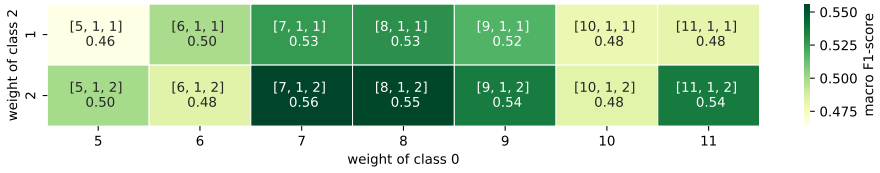



Figure 5.5: Results of the macro-F1 score on the different class weights.

Table 5.13: The results of Bert model with categorical features as text and encoded with One-Hot Encoding, where SEM stands for standard error of the mean

| | Included Features | | | Metrics | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Feature** | **Claim** | **Premise** | **Relation** | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
| **Format** | **type** | **types** | **types** | Mean | SEM | Mean | SEM | Mean | SEM | Mean | SEM |
| | x | x | x | 0.55 | ± 0.02 | 0.59 | ± 0.02 | 0.55 | ± 0.02 | 0.67 | ± 0.02 |
| **Features as text** | ✓ | x | x | 0.48 | ± 0.04 | 0.53 | ± 0.04 | 0.50 | ± 0.04 | 0.70 | ± 0.01 |
| | x | ✓ | x | **0.56** | ± **0.01** | **0.60** | ± **0.02** | **0.56** | ± **0.01** | 0.71 | ± 0.01 |
| | x | x | ✓ | 0.53 | ± 0.01 | 0.57 | ± 0.01 | 0.54 | ± 0.01 | 0.72 | ± 0.02 |
| | ✓ | ✓ | x | 0.54 | ± 0.02 | 0.58 | ± 0.02 | 0.55 | ± 0.02 | 0.71 | ± 0.01 |
| | x | ✓ | ✓ | 0.54 | ± 0.01 | 0.58 | ± 0.02 | 0.54 | ± 0.01 | 0.70 | ± 0.01 |
| | ✓ | x | ✓ | 0.48 | ± 0.05 | 0.54 | ± 0.04 | 0.49 | ± 0.04 | 0.70 | ± 0.02 |
| | ✓ | ✓ | ✓ | 0.51 | ± 0.02 | 0.55 | ± 0.02 | 0.53 | ± 0.02 | **0.74** | ± **0.01** |
| **One-Hot Encoding** | ✓ | x | x | 0.56 | ± 0.01 | 0.60 | ± 0.02 | 0.57 | ± 0.01 | 0.71 | ± 0.01 |
| | x | ✓ | x | **0.61** | ± **0.02** | **0.63** | ± **0.02** | **0.61** | ± **0.02** | **0.74** | ± **0.01** |
| | x | x | ✓ | 0.56 | ± 0.01 | 0.61 | ± 0.01 | 0.57 | ± 0.01 | 0.69 | ± 0.00 |
| | ✓ | ✓ | x | 0.58 | ± 0.01 | 0.62 | ± 0.02 | 0.58 | ± 0.01 | 0.72 | ± 0.01 |
| | x | ✓ | ✓ | 0.54 | ± 0.01 | 0.58 | ± 0.02 | 0.57 | ± 0.01 | 0.71 | ± 0.02 |
| | ✓ | x | ✓ | 0.56 | ± 0.01 | 0.60 | ± 0.02 | 0.57 | ± 0.01 | 0.71 | ± 0.01 |
| | ✓ | ✓ | ✓ | 0.55 | ± 0.01 | 0.58 | ± 0.01 | 0.55 | ± 0.01 | 0.70 | ± 0.01 |

observe a high decline in performance (49%) when including claim type and relation types.

For One-Hot encoded features, the highest macro-F1 score (61%) is achieved by including the premise types. The worst results are obtained for Bert with all features (claim/premise types and relation types). Aside from that, for all remaining sets of features, the model performs better than the model without features by at least 2%.

Overall, both approaches achieve the highest macro-F1 score when only premise type is included, which is logical in terms of argument strength assessment. Moreover, adding the categorical features either as a text or as One-Hot encoded vectors enhances the performance, in regard to F1-score, by 1% and 6%, respectively.

As a conclusion, the best model outcome is acquired using Bert with encoded categorical features with a macro-F1 score equal to 61% ± 2% outperforming the Baseline by 13% (cf. Table 5.12).

### 5.5.4 *Discussion*

Our findings confirm that it is more efficient to include the features as a One-Hot encoded vector than to include them as text. Not surprisingly, the claim and relation types are not useful to predicting the strength of an argument. This was confirmed by both the correlation matrix (cf. Figure 5.3), and the feature selection experiments (cf. Table 5.13). In contrary, the class-weight

and the new special tokens contributed positively to the learning process. Our final model architecture (cf. Figure 5.4), which outperforms its peer with no features, suggest that there is still a space for features contribution in deep learning models, as we have also seen in Chapter 3.

Finally, we calculated Pearson's correlation [Sed12] to compare our results with the recent study of [Gre+20] on the IBM-Rank-30k corpus, which they created. In their approach, the highest Pearson's coefficient corresponds to 0.52 using fine-tuned Bert while considering the topic during the training process (The topic was concatenated to the argument, using the [SEP] delimiter).

Similarly, in our case, the highest Pearson's coefficient score is achieved using Bert with encoded categorical features with a score of $0.56 \pm 0.04$. Therefore, we believe that our obtained results are consistent with the literature.

Nevertheless, this shows that language representation models still have challenges to be applied to computational argumentation [Lau21]. Moens [Moe18] extensively discussed and showed how understanding arguments in a freely uttered language is an extra difficult task for machines.

## 5.6 CORRELATION BETWEEN MANAGERS' ARGUMENTS AND ANALYSTS' RECOMMENDATIONS

In this section, we want to examine the association (if any) between the updates of analysts' recommendations and the company management's responses to analysts questions during the Q&A session in ECCs.

Actually, with earnings calls, investors quickly receive the information they want without having to search through dozens of pages in different reports. Moreover, they often schedule trades close to the earnings call, and how they trade is dependent on the information released.

However, a better (long-term) investment decisions, require other parts of fundamental analysis that are not related to the company itself. For example, to evaluate the current state of the economy and the industry of the stock sector. Therefore, many investors rely on analysts recommendations and interpretations of the market. Hence, and as we have discussed in Section 2.5.1.2, professional analysts have their observed impact on the market.

To tackle this use case, we have got analysts' announced recommendations in our studied period (2015-2019) through our collaboration with the National Institute of Advanced Industrial Science and Technology in Japan [5]. This data covers the total number of Buy, Hold and Sell recommendations, on a weekly basis. In relation to Figure 2.8, this means that recommendations are counted as follows: • *Buy* if an analyst gives a company a 1 or 2.
• *Hold* if an analyst gives a company a 3.
• *Sell* if an analyst gives a company a 4 or 5.
The reason behind this formulation is that strong buy (1) and strong sell (5) are very rarely assigned in reality [KS19]. In addition, some weeks report no new declared recommendations, and hence we have some $(N/A)$ values in our data sheet. Figure 5.6 displays the distribution of the data for our studied

5 *https://www.aist.go.jp/index_en.html*

companies (Microsoft, Apple, Facebook, and Amazon). Apparently, the Buy recommendations compose the majority, while the Sell recommendations are the minority. This is similar to the recommendations' distribution of Keith et al. [KS19] data who have 4.5% Sell (bearish), 35.7% Hold (neutral), and 59.7% Buy (bullish)[6].

To analyze the correlations, we have to point that understanding the impact of ECCs is not by the count of recommendations itself, rather by how it has been changed. Considering the recommendations themselves, we had a very high correlation, since we have in majority good quality arguments and good recommendations/price targets. To study the change in the opinions in relation to the arguments, we consider a time window of [-15,+15] days around the date of the call. Meaning that, we consider the change in recommendation is related to the ECCs itself if it happens by maximum 15 days after the call. Otherwise, the possibility that this change is related to other factors is bigger and this is not relevant for us. Similarly, we compare this change to the announced recommendations as far as 15 days before the call. In addition, we consider the change in the *ratio of Buy to Sell*, rather than the counts themselves, since this ratio reflects the change in analysts opinions more broadly.

Let $n$, $c$ be the time window limit (15 in our case), date of the call, respectively:

$$Ratio\_of\_Post\_Reco = R_{post} = \frac{Buy_{post}}{Sell_{post}} = \frac{\sum_{d=c}^{c+n} Buy_d}{\sum_{d=c}^{c+n} Sell_d} \tag{5.1}$$

Similarly,

$$Ratio\_of\_Pre\_Reco = R_{pre} = \frac{Buy_{pre}}{Sell_{pre}} = \frac{\sum_{d=\text{-n}}^{c} Buy_d}{\sum_{d=\text{-n}}^{c} Sell_d} \tag{5.2}$$

Hence, the difference in ratio (buy to sell) is:

$$Difference\_in\_ratio = D_{ratio} = R_{post} - R_{pre} \tag{5.3}$$

In addition, considering the Hold recommendations, we define the *difference_of_ratios_buy_to_all* as follows:

$$D\_buy\_to\_all = \frac{Buy_{post}}{Buy_{post}+Hold_{post}+Sell_{post}} - \frac{Buy_{pre}}{Buy_{pre}+Hold_{pre}+Sell_{pre}} \tag{5.4}$$

Likewise, we calculate the *difference_of_ratios_hold_to_all*, and the *difference_of_ratios_sell_to_all*. On the other side, we have to deal with multiple arguments stated in each targeted ECCs. For each quality metric at the level of argument (cf. Table 5.8), we consider the average value of its scores across existent arguments in this call.

Let c be one earnings call,

Let A be the total number of arguments in c,

Let M be the examined argument quality metric (e.g., argument strength):

$$M_c = \frac{1}{A}\sum_{\text{a=1}}^{A} M(\text{a}) \tag{5.5}$$

6 However, authors do not provide their data publicly, so we could not apply further comparisons.

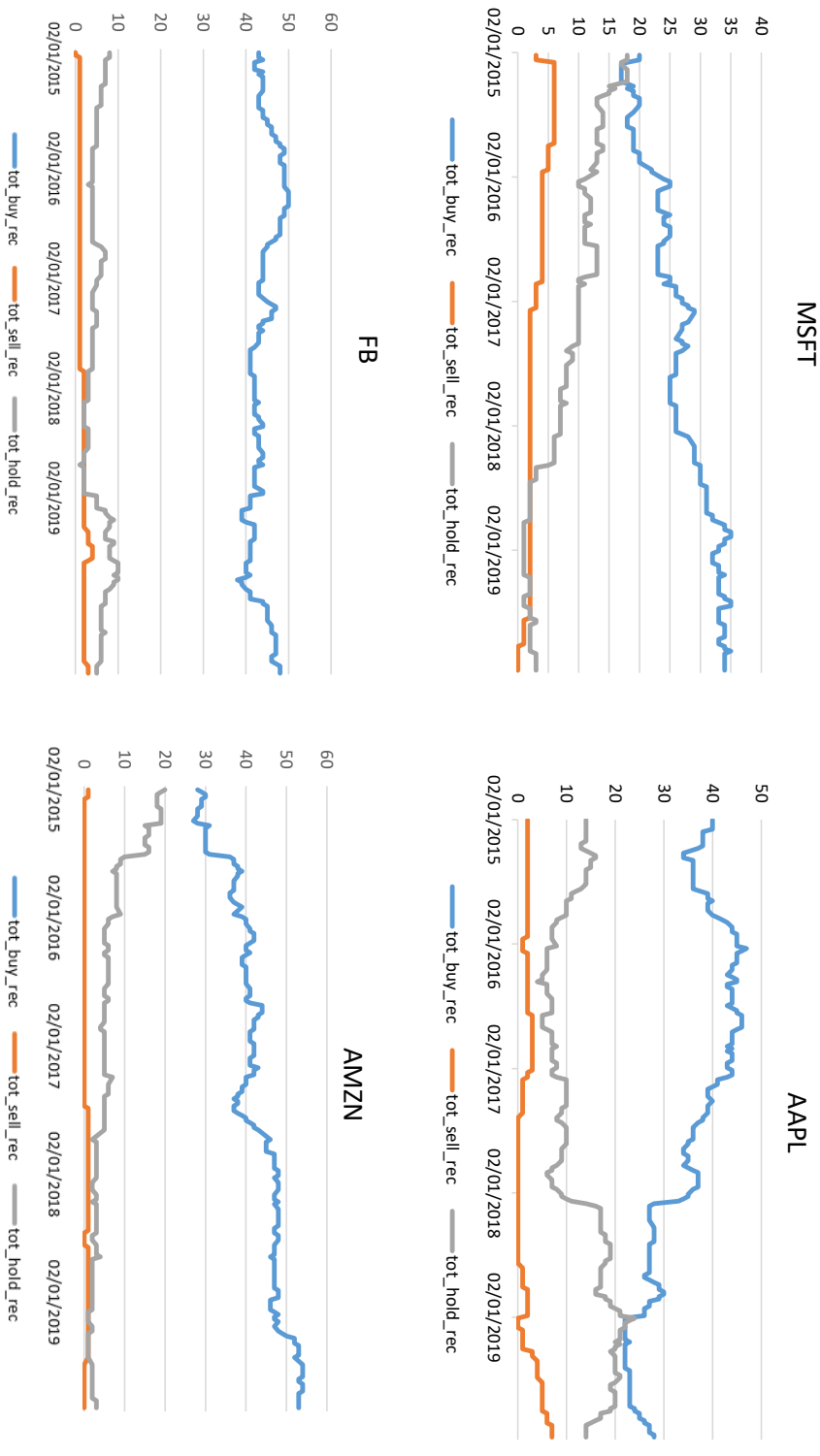



Figure 5.6: Analysts' recommendations on a weekly basis, for the studied period 2015-2019

Likewise, for argumentative unit types (e.g., premise-fact), we consider the average of count across all the call c. We exclude for this correlation the types of premise and claim that showed low frequency (e.g., claim-policy).

Figure 5.7 exhibits the found Pearson's correlation between our two variables for the four investigated companies. We can see that there are some clear correlations between our corpus features and the change ratios of analysts recommendations. For instance, argument objectivity reports 0.6 and 0.8 correlations with the *ratio_of_post_rec* [7] for Facebook and Microsoft, respectively. Yet, there is no consistent observation across all the companies. On the one hand, this could be due to the small amount of data, the disparity between companies, or the unevenness in the quality of annotators. On the other hand, this could simply represent the reality where analysts recommendations are not completely dependent on the information of ECCs.

In the recent study of Basu et al. [BX22], they attempt to answer the question "*Why donot analysts always value earnings conference calls?"*. They suggest that some analysts have some private communication channels with company insiders. Consequently, they have preempted knowledge than the information disclosure during the call. Another facet of the problem is that arguments formulate part of the manager's expressed opinion. Thus, we should consider the social judgment theory and the Friedkin-Johnsen model of opinion [FJ90]. That is, expressed opinion changes while innate opinion stays fixed. Not surprisingly, managers have to "advertise" their companies. This leads to the third facet, that good news may not always lead to bullish recommendations (cf. Section 2.5.2.2). This is also related to the fact that, what matters is not a certain metric but how it has been *changed* over the time, and whether the goal of the company is reached or not. If the company reported better-than-expected earnings, analysts' buy recommendations will rise up, most probably. If the performance is good, but still less than expectations, the recommendations as well as the price may drop. This is an important issue in the stock market, and could be detected the best in technical indicator-based studies. This is one reason which interprets their outperformance over textual analysis-based studies (cf. Figure 2.12). Yet, we have to admit that our data is not big enough and the studied companies deemed to be stable across the examined period. Unfortunately, extending to more companies was out of our funding resources. A future work must cover more diversity in the firms. For example, a small public business and maybe a bankrupt (or has been acquired) company. Finally, earnings call is just one primary piece of information that analysts (investors) consider to make recommendations (investment decisions), and it's essential to use a combination of data sources to get a better understanding of the company and the market.

## 5.7 CONCLUSION

We have introduced, in this chapter, the problem of assessing the managers' argument quality stated in their answers during earnings calls.

7 Equation 5.1.

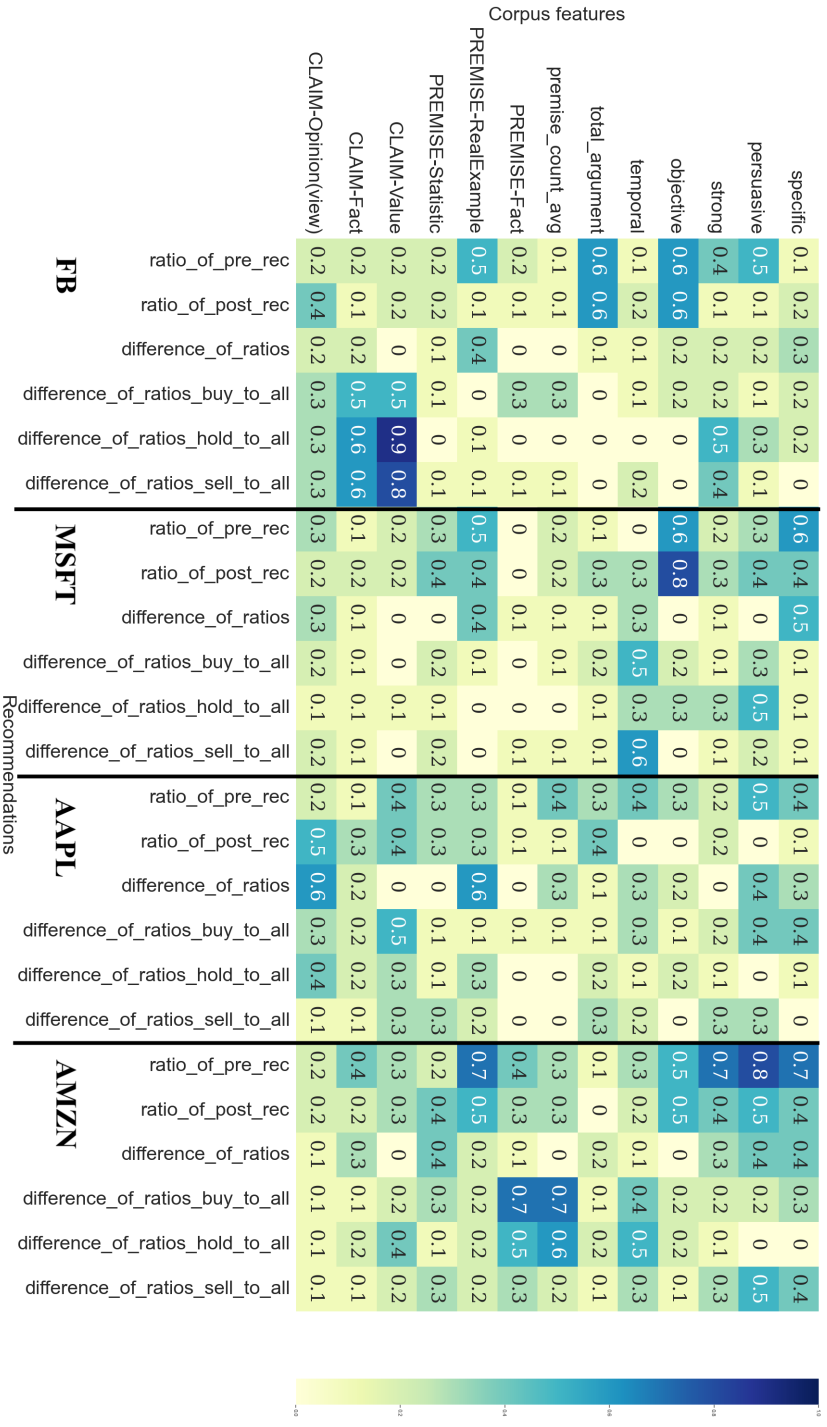



Figure 5.7: Correlation between the *FinArgQuality* and the analysts' recommendations change during 15 days.

Our argument quality metrics bridge the gap between the argumentation and financial perspectives. We hope, therefore, that our dataset *FinArgQuality* fuels further research in this interdisciplinary field.

The entire annotation process, with the frequent error checks and corrections, required about 9 months of work. Based on all our observations, we suggest addressing modal verbs, uncertainty quantification, and background-premise type in the annotation guidelines, which could further improve extended annotations. In addition, having an initial step of automatic text segmentation would increase annotators agreements on the sentence boundaries. Hence, this would improve the IAA and data quality. In addition, we suggest covering more diverse size of companies to better reflect the correlation's analysis.

The proposed model for classification of the argument Strong quality attribute, can be applied for the Specific and Persuasiveness dimensions, given the resemblance between their assessment features. Only the class-weight has to be re-calculated based on the distribution of the target dimension.

To conclude this chapter, the discipline of computational financial argumentation is relatively new. The literature displays limited resources, mainly, led by the Artificial Intelligence Research Center at the National Institute of Advanced Industrial Science and Technology in Japan, focusing mostly on the *Chinese* language. In addition, there is a recent attempt to catch up on the *Russian* language [Fis+22]. Yet, their data is not always completely available for the public.

This is, somehow, the challenge for many stock market studies, where good results are presented by financial institutes (e.g., Bloomberg) who tend to not share their data (e.g., [KS19]). Given these facts, we believe that our study on the *English* language composes an important and reliable baseline for future work.

# 6

# ARGUMENT RETRIEVAL FOR ANSWERING COMPARATIVE QUESTIONS

In this chapter, we present another interesting application of argument mining, concerning decision support systems. In particular, an intelligent web search engine whose keystone is the arguments. This direction has gained momentum since the early establishment of computational argumentation field. In this regard, the Webis Group organizes a yearly argumentation retrieval event "Touché Lab at CLEF" (2020-present)[1] that consists of two independent shared tasks:

1) Argument Retrieval for Controversial Questions.

2) Argument Retrieval for Comparative Questions.

We present in this chapter our adopted approach for participation as *"Rayla Team"* in the second shared task in 2021. The main objective of this task is to help users facing some choice problems in their daily life: Given a comparative question (for instance, "Which browser is better, Internet Explorer or Firefox?"), the aim is to retrieve documents from the ClueWeb12 corpus[2], rank and sort them based on different criteria, mainly, the arguments they provide. Table 6.1 provides a complete example from the Touché task dataset [CP21; Bon+21b].

According to [Bon+22], "at least 3% of the questions submitted to search engines are comparative". However, while some comparative questions can be answered based on facts (e.g., Is the Danube river longer than the Volga?), others need a comprehensive analysis of subjective discussions (e.g., Does Germany have a better quality of life than America?). In similar questions, we need to mine the arguments shown in the expressed opinions. Bondarenko et al. [Bon+20a] mentioned that more than 65% of comparative queries demand argumentation. Developing such a technology is precisely the target of this chapter.

In order to have more granularity control, our architecture incorporates several, yet complementary, units. Each one is dedicated to perform a specific sub-task. Namely, query expansion, argument mining, scoring, and sorting.

The Touché organizers offer the participants to use their own TARGER [Che+19] tool, for the argument retrieval sub-task. Since this is the main engine for our target model, we decided to use our own module based on the latest developments in the field of computational argument mining. Therefore, we

1 https://webis.de/events/touche-21/index.html

2 https://lemurproject.org/clueweb12/

Table 6.1: Example of query and document with different relevance in Touché task dataset, Source: [CP21; Bon+21b]

| Query | Document | Rank |
|---|---|---|
| What is better for the environment, a real or a fake Christmas tree? | "Disease and condition content is reviewed by our medical review board real or artificial? There is so much confusing information out there about which is better for your health and the environment." | 2 |
| | "You may think you're saving a tree, but the plastic alternative has problems too. Which is "greener" an artificial Christmas tree or a real one? " | 1 |
| | "This entry is part 25 of 103 in the series eco-friendly friday november 28th's tip christmas trees: stuck between choosing a real Christmas tree or a fake one?" | 0 |

implemented, for our participation, a new transfer learning model for argument identification based on DistilBERT [San+19]. In addition, we aim at testing the real impact of improving the argument mining model itself. Hence, we injected our ensemble learning model [Alh+21b] instead of DistilBERT in the same submitted global architecture. We subsequently test its outcome on the shared task data. Although our point of interest is argument mining, we need to retrieve a good set of relevant documents at first place (to be further inspected for arguments). This depends basically on the request query. Furthermore, the final selection and sort of documents shown to the user is in function of the ranking algorithm. Thus, every unit in this model has its impact on the end result. We, therefore, extend upon each unit in Section 6.2 before moving to the evaluation outcomes in Section 6.3. We conclude our findings and future directions in Section 6.4.

## 6.1 RELATED WORK

Question answering is a sophisticated form of information retrieval that started to develop as a field decades ago. However, with the growth of World Wide Web data, as well as of private databases, the need for more precise, well-expressed and shortly formulated answers is growing, too. This could be defined as Focused Retrieval (FR) which deals with retrieving specific information [JSAH07]. Hence, several studies are devoted to the representation of natural language stated in the query and in the documents.

Extracting the arguments declared in the document is one way to clearly capture the grounded statements (premises) and the final conclusion (claim) given in the text. Therefore, many recent works focus on the arguments as a potential tool for improving comparative question answering [Pan+18; Sch+19; Bon+20b], and more generally, on building an argument-based search engine as in the work of Daxenberger et al. [Dax+20] with respect to their *summetix*

project (formely known as ArgumenText)[3]. Similarly, the *args.me* project[4] [Ajj+18; Ajj+19].

The most relevant to this work is the previous shared task Touché 2020 [Bon+21a]. The best result of that shared task was introduced by the team "Bilbo Baggins" of Abye et al. [AST20]. They used a two-step approach: 1) query expansion by antonyms and synonyms, 2) document ranking based on three measures: relevance, credibility, and support features. An overall score is then drawn by summing up those scores after weights multiplication. In [Huc20], Huck et al. participated with a simple approach where the original topic is used as a single query to retrieve the top 20 documents. They then used the TARGER API [Che+19] in order to extract argumentative text from each document. For re-ranking, the BM25 algorithm (also known as *Best Matching 25*) has been used to determine the relevance of the extracted arguments with respect to the original query. With respect to 2021 edition, the "Katana" team [CP21] scored similar results to ours, according to the official metrics of the competition (nDCG@5). They keep the original questions as queries and re-rank the top-100 retrieved results using different models. However, their best performance was obtained by gradient boosting methods, training on ranking cost function: XGBoost and LightGBM.

Our approach considers also the query expansion and multiple scoring criteria. However, instead of using a static classifier 'XGBoost' , we build on a transformer model which can ameliorate the robustness of the classifier and extend its operating range.

## 6.2 METHOD

In this section, we present our proposed approach and adopted methods to build a search engine for answering comparative questions based on argumentation identification. The overall architecture of our approach is presented in Figure 6.1. It consists of a sequence of seven stages. We extend on them individually in the upcoming sections. We used the same architecture to submit four runs with different configurations via TIRA platform [Pot+19].

### 6.2.1 *Query Expansion*

Query Expansion (QE) is the process of reformulating a given query to improve the retrieval performance and increase the recall of the candidate retrieval [AD19]. Our query expansion module involves a variety of techniques in order to generate three different queries to be passed to the next step, as the following:

- *Query 1:* is the original query itself.
- *Query 2:* focuses on the comparison objects. It is generated from the original query by: (1) removing English stop words, punctuation marks,

3 https://www.summetix.com/

4 https://www.args.me/index.html

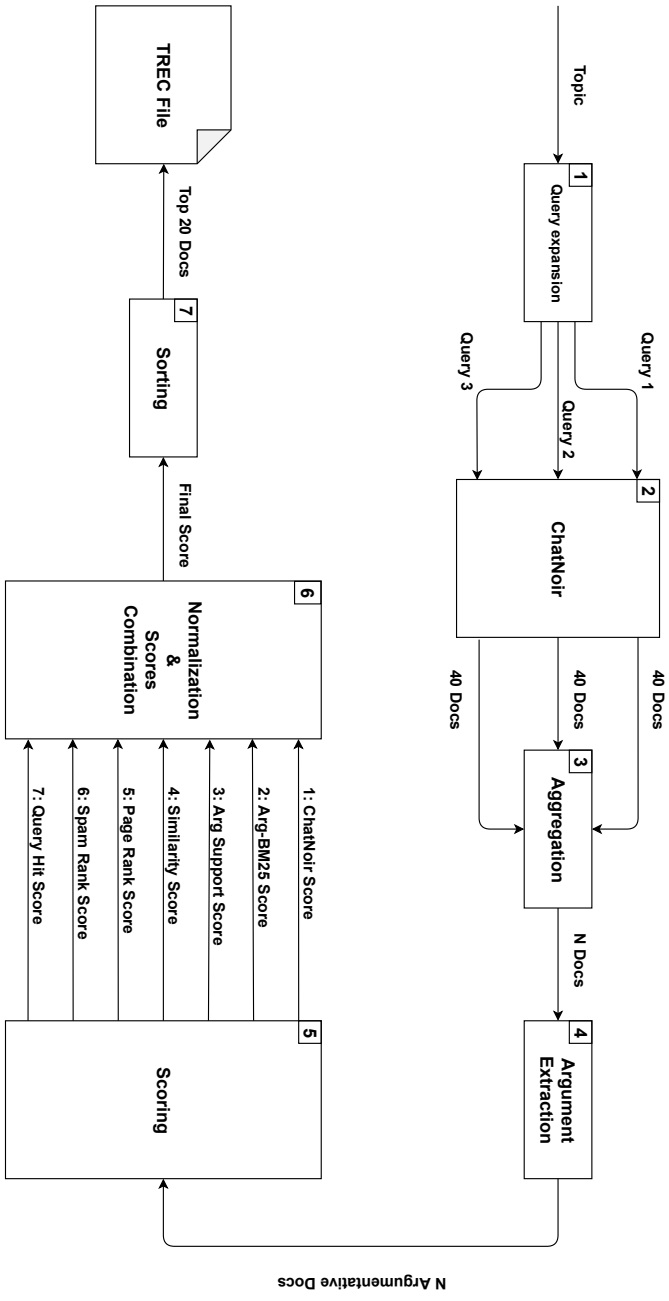



Figure 6.1: Global architecture of the submitted approach

and comparison adjectives also called comparison operators. (2) Stemming of the remaining words to their base forms, and aggregating them together with conjunctive AND operator.

- *Query 3:* focuses on the comparison aspect. Hence, it will be generated from the original query only if the latter contains a comparison operator, as follows:
  - Search for synonyms and/or antonyms of the comparison operator of the query to get the context of the comparison operator. We consider five synonyms/antonyms in our case.
  - Remove English stop words and punctuation marks.
  - Eliminate the comparison operator from the original query and stemming the remaining words/terms to their base form.
  - Create 5 queries out of the original query by adding one of the synonyms/antonyms extracted earlier. Those 5 output queries are sent to ChatNoir API as one disjunctive OR-query: Query 3.

Comparison adjectives identification and words stemming are done automatically using SpaCy[5]. Synonyms and antonyms are hard-coded in the software due to the limited number of comparative operators in the topics of both years' data (10 adjectives in total).

Table 6.2 shows an example of the query expansion output for the Topic 61 from Touché 2021 topics [Alh+21a].

Table 6.2: Example of Query Expansion

| Query id | Generated query |
|---|---|
| Query 1 | Who is stronger, Hulk or Superman? |
| Query 2 | Hulk *AND* Superman |
| Query 3 | Hulk *AND* Superman *AND* strong *OR*<br>Hulk *AND* Superman *AND* capable *OR*<br>Hulk *AND* Superman *AND* powerful *OR*<br>Hulk AND Superman *AND* able *OR*<br>Hulk *AND* Superman *AND* weak |

### 6.2.2 *Document Retrieval by ChatNoir API*

The ClueWeb12 document dataset for this task is easily accessible through the ChatNoir API [6] [Bev+18; Pot+12] that is based on the BM25F ranking algorithm [RZT04]. The BM25 (Best Matching 25) is a popular ranking method

5 https://spacy.io/

6 https://www.chatnoir.eu/doc/api/

used by several search engines to determine the relevance of documents in regard to a query. It is based on the probabilistic retrieval framework, and it is considered as the state-of-the-art among TF-IDF-like algorithms in information retrieval. BM25F is just a newer version of BM25 that takes the structure of documents into consideration. The API takes a query as input, and returns a set of documents as specified. In our experiments, we have tested the system with several numbers of document counts ($M$ that is a parameter of the system) for each of the three queries [7].

### 6.2.3 *Document Aggregation*

As Query 3 may be empty, we have a minimum of $2 \times M$ and a maximum of $3 \times M$ retrieved documents ($M$: retrieved docs per query). Every document has a unique id (uuid), which we use to remove the redundant documents returned by more than one query. For instance, if a document is retrieved by Query 1 and Query 2, with different ChatNoir scores for each (*score*, *page_rank* and *spam_rank*), the document aggregation component will output one document with scores deduced by the sum of scores from Q1 and Q2 ($score\ \ score_1\ score_2$, ...).

Initially, the ChatNoir API does not respond with the full content of the documents. Instead, it returns some metadata, including the unique uuid. To get a document full content, a new query is sent to ChatNoir to request the full HTML source page. Later on, we need to reveal only the main textual content from the HTML document. We achieve that by removing tags, advertisements through a cleaning process. For that end, we used two HTML cleaning libraries: *Boilerpy3* [8] and *Trafilatura* [9].

### 6.2.4 *Argument Extraction*

In our particular task, we seek to detect the comparative sentences in the document, therefore, argument identification can be sufficient (i.e., detecting of argumentative text). Hence, we take the sentences from the document aggregation step and apply binary classification using either DistilBERT (cf. Figure 3.2) or the stacked model (cf. Figure 3.3) to label every sentence as an argument or non-argument.

We have used the same corpora: Student Essays [SG14a] and Web discourse [HG17] for training. Therefore, you may have a look at the detailed results in Table 3.2, Table 3.3 and Table 3.4.

---

7 The interaction with this API can be done using either GET or POST requests, several parameters could be included in the request to specify the corpus to retrieve from, and the type of returned information. It supports also a set of standard operators from web search services, and queries can be concatenated using *Boolean operators, "-", "... ", site:... etc.*. In our approach, we used only POST requests, *AND* and *OR* operators to concatenate strings in Query 2 and Query 3.

8 https://github.com/jmriebold/BoilerPy3

9 https://github.com/adbar/trafilatura

### 6.2.5 *Scoring*

The scoring or ranking step is essential for any search engine system because many users tend to check out the top results without spending time to carefully review the later ones. Subsequently, our objective now is to estimate the best matching between the query and the candidate answers, in order to sort them at the final stage. To this end, we investigate different scores based on different aspects. Foremost, the document relevance, which can be checked simply by ChatNoir BM25 score and Query hit count.

However, even if a document content is relevant to the query, it may be fake or biased. Thus, we inspect the credibility of the document itself by considering: Page Rank score as well as Spam rank score, we will detail in the following.

Moreover, as we built our retrieval system based on arguments, we take into consideration the argument quality level by three different scores: argument support, query-argument similarity, and argument BM25 score.

We refer to each of our ranking scores by a score-id from (1) to (7) to be further used in Table 6.3. The complete details of those scores are addressed in the following:

- (1) ChatNoir score: returned form ChatNoir API indicating BM25 measure.
- (2) Arg-BM25 score: calculated on argumentative sentences of each document with respect to the original query. This is done through re-indexing the retrieved documents by creating new ones that contain only argumentative sentences. Then the arg-BM25 score of each document is calculated by querying the new argumentative documents with the original topic.
- (3) Argument support score: Assuming that the more arguments the document has, the more interesting it is, we here pay attention to the ratio of argumentative to all sentences in the document.
- (4) Similarity score: evaluates the similarity of two sentences based on the context and English language understanding using the *Sentence-Transformer* [10] library [RG19]. We calculate the similarity between the original query and every argumentative sentence in the document, and consider the average.
- (5) Page Rank score: given by ChatNoir API, measuring the importance of the source website pages.
- (6) Spam rank score: given by ChatNoir API, indicating the probability of the website to be a spam.
- (7) Query hit count: indicates how many times the document is retrieved by the three queries [1,3].

10 https://github.com/UKPLab/sentence-transformers

Table 6.3: Configurations of each run: scores are defined in Section 6.2.5 with respect to the score-ids (1) to (7)

| | Score Weights | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Run Tag | (1) | (2) | (3) | (4) | (5) | (6) | (7) | Docs (M) |
| DistilBERT_argumentation_bm25 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 30 |
| DistilBERT_argumentation_advanced_ranking_r1 | 15 | 25 | 25 | 15 | 20 | 0 | 0 | 20 |
| DistilBERT_argumentation_advanced_ranking_r2 | 10 | 10 | 50 | 20 | 10 | 5 | 5 | 40 |
| DistilBERT_argumentation_advanced_ranking_r3 | 10 | 15 | 10 | 50 | 10 | 0 | 0 | 40 |
| Stack_argumentation_bm25 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 30 |

Table 6.4: Results of each run. NDCG (Normalized Discounted Cumulative Gain)

| | Relevance | | Quality | |
|---|---|---|---|---|
| Run Tag | NDCG@5 | Rank/20 | NDCG@5 | Rank/20 |
| DistilBERT_argumentation_bm25 | 0.466 | 6 | **0.688** | 1 |
| DistilBERT_argumentation_advanced_ranking_r1 | **0.473** | 3 | 0.670 | 5 |
| DistilBERT_argumentation_advanced_ranking_r2 | 0.458 | 8 | 0.630 | 11 |
| DistilBERT_argumentation_advanced_ranking_r3 | 0.471 | 4 | 0.625 | 13 |
| Stack_argumentation_bm25 | 0.444 | *N* | 0.640 | *N* |
| Touché baseline | 0.422 | 6 | 0.636 | 6 |

### 6.2.6 *Normalization and Scores Combination*

For the final score, we normalize all previously calculated scores, so that all values are between 0 and 1. These scores are aggregated using particular weights, which we set up experimentally based on the announced relevance judgments of the 2020-Touché task 2.

### 6.2.7 *Sorting*

At this stage, the documents are sorted based on the final score to get the top 20 documents that are highly relevant to answer the comparative query. The final output is inserted into a text file while respecting the standard TREC format proposed by the Touché organizers.

## 6.3 TOUCHÉ SHARED TASK EVALUATION

The shown architecture in Figure 6.1 represents our base approach, from which we derive four submissions to the task-2 of Touché 2021 by experimentally modifying score weights and the number of retrieved documents. Table 6.3 describes the different configurations.

The Touché committee, with the help of volunteers, identifies two scores for each document (with respect to the original topic/query):
1) The relevance (2: high relevant, 1: relevant, 0: not-relevant).
2) The quality: the degree of the document readability and the usage of rhetorical figures. This is in the scale 0 to 2 as well.

Based on that, the official judgments, are calculated in terms of NDCG (Normalized Discounted Cumulative Gain) score, which is commonly used by search engines. Table 6.4 shows our outcomes of each of our DistilBERT runs [Bon+21b]. Moreover, we were able to further produce the results of our system after plugging the stacking model, using the evaluation script provided later by the organizers. This belongs to *Stack_argumentation_bm25* as shown in Table 6.4.

By comparing our results with the Touché baseline, which is a BM25F algorithm, we observe that our approach outperforms the baseline in terms of relevance and quality. This indeed confirms that the problem of answering comparative questions should not be addressed as a traditional document retrieval problem.

Last but not least, we would like to point that our participation scored first place among all the teams with respect to quality and the third in terms of relevance by achieving a score of *NDCG@5* 0.688 and *NDCG@5* 0.473 respectively [Bon+21b]. Our submitted notebook paper [Alh+21a] was further nominated and published (after extension) as *«Best of 2021 Labs»* in the CLEF 2022 conference [Alh+22a].

## 6.4 CONCLUSION

Despite the wealth of counter information available on the web, it is still hard for search engines to elaborate on comparative queries by complete precision. This chapter sheds the light on the crucial role of argumentation in answering comparative questions.Every component in our procedure (cf. Figure 6.1) contributes to the end result. For instance, the query expansion component makes a first selection and build a set of topic-related documents. The three different queries generated from the original topic increase the coverage of the related documents and this works very well with the ChatNoir API since it is a basic BM25F retrieval system.

When plugging the stacked model instead of the DistilBERT in the overall architecture (*Stack_argumentation_bm25*), we observe that it did not achieve the expected improvement over DistilBERT runs. This could be due to the type of text retrieved from the ClueWeb12. In fact, the training of the SVM is based on textual features, such as 1-3 gram Bag of Words (BoW) and Named Entity Recognition (NER), which are limited to the type of the text during the training, unlike the DistilBERT model which generalize over text and grasp more contextual knowledge as we have discussed already in Chapter 3.

On the other hand, chaining seven components in one sequence may generate easily an error propagation and amplification (aka. Inheritance bias). For instance, by only increasing the initial number of retrieved documents with a very high value, the final score is negatively degraded. This is in consequence of including some unrelated documents to the succeeding components. This could explain the lower quality scores of runs *DistilBERT_argumentation_advanced_ranking_r2* and *DistilBERT_argumentation_advanced_ranking_r3*.

As any search engine, the improvement of any individual process can reflect positively on the user output. In our future work, we plan to touch particularly the following points: to improve the ranking stage by using an argument quality assessment model, as well as developing a machine learning method that learns the best weights of our ranking scores (given that we can now get the needed data from the three challenge edition' data).

Apart from this specific challenge, the good treating of comparative questions requires the right classification of the query as comparative in the first place. In addition, handling of special cases, like implicit comparison object (e.g., "Is homeschooling better?") or unclarified comparison aspect (e.g., "Is it better to live abroad?"). Working on those indirect questions can provide a better interpretation of what is the information need beyond the comparative question.

# 7

# CONCLUSION AND FUTURE WORK

This chapter concludes the work of the thesis and gives an outlook for future directions.

To conclude this work, we start by recalling the workflow as presented in this thesis. In Chapter 1, we motivated our work by revealing the value of argument mining in the financial domain, besides the limitations of existing literature. We have also discussed the historical dimension of argumentation theory, which led to the numerous proposals of argument models in the first place, and of argument quality assessment criteria in the second place. In Chapter 2, we reviewed the state-of-the-art over those axes of research, and their interplay through FinNLP and argumentation in finance.

In Chapter 3, we detailed further on the related work in terms of the robustness of argument mining domain generalization, and we present our intensive study towards it. Later on, in Chapter 4 and Chapter 5, we introduced our contributions towards **FinArg** and **FinArgQuality**, and their correlations with the recommendations of professional analysts. Finally, in Chapter 6, we showed another use case of argument-based decision support systems, in a real-world problem of answering comparative questions based on the retrieved arguments.

The main contributions of this thesis are summarized, discussed and extended to the future perspectives in the following paragraphs.

### EXPERIMENTAL EVALUATION OF THE ROBUSTNESS OF ARGUMENT MINING MODEL OVER SHIFTED DOMAINS

With respect to the research question *RQ1*: "How to build and evaluate an argument parser that can generalize over heterogeneous corpora, given that argument mining is a domain-dependent task?", this thesis provides an empirical study on the robustness of two argument mining models over shifted domains and different model runs.

Our findings suggest that a stacking approach, which composes of SVM and DistilBERT evinces more stability, in most of the cases, than the fine-tuned model of DistilBERT solely. This implies some limitation for the transfer learning model of DistilBERT when dealing particularly with argument unit classification task over shifted data. In other words, there are still some text features (e.g., count features) that DistilBERT knowledge is still missing. This is in line with the research of incorporating external knowledge to pre-trained language models (e.g., [Zha+20b; Zha+20c; WPS21; CH+21]).

In addition, the stacked model showed lower variance in the standard deviation than DistilBERT in almost all the experimented cases. This behavior of transformers is also highlighted by [MML20] who compared the performance of 100 instances of BERT for both in-domain and out-of-domain tests. They demonstrated that those instances varied largely in their out-of-distribution generalization performance. In this regard, measuring the distance between two domains is till an open research question, which we intend to address in our future work.

Furthermore, in many generalization experiments, it is hard to find an explanation, since there is no one consistent pattern. Hence, future work to interpret and understand what makes a model perform better in some cases is a demanding need. One challenge could be, the increasing number of parameters the recent language models consider. For example, while the Bert model has 340 million parameters [Dev+18], the PaLM model, developed in 2022, [Cho+22] considers 540 billion parameters. Apparently, the notion of what is a *large* language model is changeable through (short) time.

However, this raises the question: to what extent can we replicate those models? Especially, in academic settings, where significant capital sources may be limited (in terms of data and infrastructure). Moreover, in terms of environment: Strubell et al. [SGM19], studied the energy consumption of deep language models. They show that the estimated $CO_2$ emission from training a single AI model can go up to an average of five cars consumption in their lifetimes [1]. They, hence, suggest that "researchers should prioritize computationally efficient hardware and algorithms".

Another limitation of many recent language models is their accessibility. The source code of some models is not provided. The access is only through an API (e.g., GPT-3 [Bro+20]), while no access is even given for some other models (e.g., LaMDA [Tho+22]). Thus, the interpretation or improvement over those models is challenging.

Finally, we believe that we should examine more unsupervised approaches in computational argumentation research. This may help to solve the data shortage in this discipline.

To sum up, the generalization of NLP models is a complex and ongoing problem, and there is much future work to be done in this area. From large-scale pre-training and domain adaptation to incorporating structural information and developing interpretability and explanation methods.

### ARGUMENT STRUCTURE IN EARNINGS CONFERENCE CALLS

With respect to the second research question (*RQ2*), this work paves the way for scholars in the FinNLP domain to use argumentation as a tangible practical instrument for financial applications.

We chose the basic form of argument structure that focuses on its mandatory components: the premise(s), the claim, and their relation(s). This choice was based on the analysis of both proposed argument models in computational

1 https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

argumentation (i.e., practice of creation of a dataset), and the normative formulation of the argument structure and argumentation scheme in argumentation theory. Hence, we modeled the structure of *each argument* using *one-claim-approach* proposed by Cohen [Coh87]. By that, we achieved an agreement of unitized Krippendorff's $\alpha_U$ =0.70 for argument components and Krippendorff's $\alpha$=0.81 for argument relations. Hence, we conclude that the annotation of arguments in ECCs is reliably possible.

We have evaluated the automatic argument identification, and argument unit classification tasks on *FinArg*, using the stacked model proposed earlier. We have achieved a high level of effectiveness for both tasks.

In addition, we view the entire answer of a company representative as one chain of reasoning, which can contain multiple arguments. However, we did not consider the external relations between arguments. This could be one direction for future work, in order to structure the complete dialogue.

### ARGUMENT QUALITY ASSESSMENT IN EARNINGS CONFERENCE CALLS

There is no doubt that the judgment of the quality of an argument is bilateral to the acceptance of its method of inference and to the application-specific variants.

With respect to *RQ3*, we aimed, in this contribution, to establish a well-considered link between insights derived from practitioners in CAQ, and the literature of financial text analysis. Nevertheless, we first reviewed and studied different proposals of argument quality assessment in argumentation theory such as fallacies, and RAS criteria (see Section 2.4).

Hence, we have defined five metrics of overall argument quality, as well as types of common observed claims and premises in ECCs. We introduce *FinArgQuality*, the first financial corpus annotated with argument quality.

Moreover, using our data, we have examined the multi-class classification task of argument strength on a scale zero to two. To this end, we developed a refined version of BERT, that integrates some categorical features with one-hot encoding. This outperforms the BERT baseline by $13 \pm 2\%$ in regard to F1-score. Our study confirms two points. First, the type of premises shows a higher impact on the overall argument strength than the claim type. Second, despite the competent understanding capability that BERT shows in a wide range of studies and benchmarks, the estimation of the argument quality is still a challenging task. This is because arguments are an advanced form of human language utterance where some entailment elements can be implicit, and external domain knowledge is still required to solve ambiguity.

However, we plan to study other open source language models, and other quality attributes. Additionally, we intend to conduct a hold-out experiment in terms of company (leave-one-company-out), in order to study the similarities across data sources from the same industrial sector.

However, we believe that having more financial support to get a large scale corpus, covering a wide spectrum of sectors and companies, will certainly

enable a better performance of BERT, better understanding of internal associations (within sectors), and external correlations with the market.

Through the writing of this thesis, both financial NLP and computational argumentation communities have suffered from the lack of labeled data. Therefore, we believe that our carefully developed corpora can prompt future directions.

On the one hand, Figure c.1 shows a taxonomy of six potential argument mining tasks using our data. In addition, a future direction could be to use the argumentative unit types in order to mine the argumentation strategies [AK+16b].

On the other hand, the automatic detection and qualification of arguments in financial domain is important for several goals:

- Efficiency: It allows for the analysis of large amounts of financial data 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.
- Improved decision-making: Providing a comprehensive and objective analysis of financial arguments, identifying worthiness, and detection of verified claims, can help to inform and improve decision-making in finance.
- Enhanced market transparency: Arguments can provide more visibility into the reasoning behind investment decisions or analysts' recommendations, improving market transparency and trust.

Overall, some general open challenges are:

- Ambiguity: Financial texts often contain complex language, technical terms, and specialized vocabulary.
- External and multimedia resources: The usage of external resources such as links or images in social media, charts, and tables in ECCs presentation make it harder to automatically structure the argument and to estimate its quality. This is because these forms of content are not easily processable by NLP algorithms. Yet, they may contain additional information or conflicting perspectives.
- Argument Quality: There is no consensus on the right measurement scale for argument quality, with a variance in studies (binary classification, 3, 5 points scale, or more).
- Estimating the influence period (time-window) of a financial opinion.
- Natural language understanding of argumentative messages is still challenging for machines.

As a first emerging output of this thesis, the FinArg-1 Shared Task, in cooperation with AIST, Japan [2] and different other partners has been established [3]. This task is planned for three years on both Chinese and English languages. We will cover different tasks from argument identification, relation classification, argument-based sentiment analysis, to argument quality assessment. Our ultimate goal is to improve the automatic understanding of financial text.

We hope that the work presented in this thesis fuels and inspires more research in computational argumentation, stock market and their interplay.

2 https://www.aist.go.jp/index_en.html

3 http://finarg.nlpfin.com/

Part I

# APPENDIX

# a

## ADDITIONAL EVALUATION OF THE MODEL ROBUSTNESS IN DOMAIN GENERALIZATION

We report in this section, the macro averages of precision, recall, F1-score and Accuracy, for the model robustness experiments which were done in Chapter 3:

- Multi-dataset Learning (Section 3.4).
- Cross-domain Settings: Testing on a Completely Unseen dataset (Section 3.5).
- Cross-topic Settings: Testing on Completely Unseen Topics (Section 3.6).

However, the interpretation of those results maintain unchanged, since they still lead to the same conclusions.

Table a.1: SDL vs. MDL argument identification using the stacked model, where Std stands for standard deviation and the drop is calculated compared to the SDL results of each dataset separately

| | | **Macro-Precision** | **Macro-Recall** | **Macro-F1 score** | **Accuracy** |
|---|---|---|---|---|---|
| | Dataset | Mean Std | Mean Std | Mean Std | Mean Std |
| SDL | SE | 0.900 ±0.004 | 0.831 ±0.004 | 0.864 ±0.004 | 0.920 ±0.002 |
| | WD | 0.725 ±0.015 | 0.705 ±0.015 | 0.715 ±0.020 | 0.776 ±0.011 |
| MDL | SE | 0.864 ±0.016 | 0.704 ±0.016 | 0.776 ±0.007 | 0.881 ±0.004 |
| | WD | 0.721 ±0.012 | 0.610 ±0.012 | 0.661 ±0.024 | 0.765 ±0.009 |

Table a.2: SDL vs. MDL argument unit classification using the stacked model, where Std stands for standard deviation and the drop is calculated compared to the SDL results of each dataset separately

| | | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|
| | Dataset | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SDL | SE | 0.815 | ±0.002 | 0.801 | ±0.002 | 0.808 | ±0.004 | 0.827 | ±0.003 |
| | WD | 0.790 | ±0.019 | 0.816 | ±0.019 | 0.803 | ±0.016 | 0.868 | ±0.010 |
| | IBM | 0.987 | ±0.002 | 0.987 | ±0.002 | 0.987 | ±0.002 | 0.987 | ±0.002 |
| MDL | SE | 0.727 | ±0.153 | 0.662 | ±0.153 | 0.693 | ±0.141 | 0.738 | ±0.125 |
| | WD | 0.673 | ±0.024 | 0.667 | ±0.024 | 0.670 | ±0.035 | 0.796 | ±0.015 |
| | IBM | 0.910 | ±0.006 | 0.879 | ±0.006 | 0.894 | ±0.008 | 0.895 | ±0.008 |

Table a.3: Evaluation of the cross-domain argument identification task, where Std stands for standard deviation.

| | | | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Training | Testing | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SE | WD | stacked model | 0.469 | ±0.006 | 0.407 | ±0.006 | 0.436 | ±0.009 | 0.455 | ±0.013 |
| | | DistilBert | 0.596 | ±0.004 | 0.548 | ±0.004 | 0.571 | ±0.005 | 0.694 | ±0.003 |
| | | [MS16] | | | | | 0.524 | | 0.524 | |
| WD | SE | stacked model | 0.618 | ±0.013 | 0.581 | ±0.013 | 0.599 | ±0.009 | 0.771 | ±0.012 |
| | | DistilBert | 0.657 | ±0.016 | 0.519 | ±0.016 | 0.580 | ±0.015 | 0.798 | ±0.005 |
| | | [MS16] | | | | | 0.128 | | 0.181 | |

Table a.4: Evaluation of the cross-domain argument unit classification task, where Std stands for standard deviation.

| | | | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Training | Testing | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| SE, WD | IBM | stacked model | 0.759 | ±0.015 | 0.436 | ±0.015 | 0.554 | ±0.079 | 0.61 | ±0.052 |
| | | DistilBert | 0.698 | ±0.027 | 0.353 | ±0.027 | 0.469 | ±0.023 | 0.55 | ±0.013 |
| SE, IBM | WD | stacked model | 0.590 | ±0.116 | 0.370 | ±0.116 | 0.455 | ±0.196 | 0.546 | ±0.281 |
| | | DistilBert | 0.662 | ±0.023 | 0.551 | ±0.023 | 0.602 | ±0.012 | 0.805 | ±0.009 |
| WD, IBM | SE | stacked model | 0.678 | ±0.012 | 0.429 | ±0.012 | 0.526 | ±0.060 | 0.675 | ±0.016 |
| | | DistilBert | 0.293 | ±0.064 | 0.487 | ±0.064 | 0.366 | ±0.057 | 0.586 | ±0.128 |

Table a.5: Model assessment in cross-topic experiments for argument identification task. |S|: number of Sentences/Topic, |T|: number of Topics, Std: standard deviation

| | | | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| \|S\| | \|T\| | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 4 | 300 | stacked model | 0.859 | ±0.038 | 0.759 | ±0.038 | 0.806 | ±0.029 | 0.895 | ±0.019 |
| | | DistilBert | 0.643 | ±0.176 | 0.505 | ±0.176 | 0.566 | ±0.095 | 0.825 | ±0.030 |
| 6 | 200 | stacked model | 0.820 | ±0.028 | 0.719 | ±0.028 | 0.766 | ±0.017 | 0.862 | ±0.008 |
| | | DistilBert | 0.583 | ±0.184 | 0.418 | ±0.184 | 0.487 | ±0.056 | 0.791 | ±0.021 |
| 24 | 50 | stacked model | 0.778 | ±0.057 | 0.573 | ±0.057 | 0.660 | ±0.038 | 0.813 | ±0.026 |
| | | DistilBert | 0.447 | ±0.169 | 0.431 | ±0.169 | 0.439 | ±0.012 | 0.775 | ±0.030 |

Table a.6: Model assessment in cross-topic experiments for argument unit classification task. |S|: number of Sentences/Topic, |T|: number of Topics, Std: standard deviation

| | | | **Macro-Precision** | | **Macro-Recall** | | **Macro-F1 score** | | **Accuracy** | |
|---|---|---|---|---|---|---|---|---|---|---|
| \|S\| | \|T\| | Model | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 3 | 400 | stacked model | 0.794 | ±0.020 | 0.780 | ±0.020 | 0.787 | ±0.015 | 0.800 | ±0.014 |
| | | DistilBert | 0.757 | ±0.025 | 0.759 | ±0.025 | 0.758 | ±0.025 | 0.767 | ±0.023 |
| 4 | 300 | stacked model | 0.819 | ±0.028 | 0.790 | ±0.028 | 0.804 | ±0.036 | 0.820 | ±0.032 |
| | | DistilBert | 0.756 | ±0.031 | 0.740 | ±0.031 | 0.748 | ±0.031 | 0.766 | ±0.031 |
| 6 | 200 | stacked model | 0.820 | ±0.018 | 0.814 | ±0.018 | 0.817 | ±0.020 | 0.825 | ±0.019 |
| | | DistilBert | 0.780 | ±0.018 | 0.778 | ±0.018 | 0.779 | ±0.018 | 0.786 | ±0.019 |
| 24 | 50 | stacked model | 0.851 | ±0.068 | 0.841 | ±0.068 | 0.846 | ±0.070 | 0.855 | ±0.055 |
| | | DistilBert | 0.831 | ±0.093 | 0.825 | ±0.093 | 0.828 | ±0.091 | 0.835 | ±0.079 |

# b

# GUIDELINES FOR ANNOTATING ARGUMENT STRUCTURE

## B.1 INTRODUCTION

Our dataset consists of the earnings calls of four S&P 100 stock market companies over the five years from 2015 to 2019. An earnings call is a quarterly organized event between a company's management and professional analysts (possibly also investors and business journalists). The call starts by the company reporting their financial performance in the last fiscal quarter and giving their estimations for the upcoming one. At the end of each earnings call, there is a Q&A session where analysts are allowed to ask questions about the presented information and discuss their analysis and views. The analysts, later on, announce (update or maintain) their recommendations to the public which could be a reference for investors who make their trades, accordingly.

However, with the goal of consensus or persuasiveness, managers tend to use arguments in their answers.

The argument is the logical reasoning and justification stated to achieve a final conclusion. The minimum argument consists of: one premise (also called an evidence/reason) supporting one claim (also known as conclusion).

The task of the annotator is to: (1) separate argumentative from non-argumentative parts within the company's representative answers, (2) identify the different argument components and link them using support (or attack) relations. (3) In addition, to define the quality of the arguments. We clarify this subtask in details in a separate file (cf. Appendix c).

## B.2 ANNOTATION PROCESS

### B.2.1 *Overview*

Generally, the replica of the analyst is not argumentative, but the questions give us a picture of the context and the topics to be discussed. Therefore, questions have to be carefully considered, which is also important to determine the argument quality. An argument can be seen as a combination of a conclusion (in terms of a claim) and a set of premises (in terms of supporting reasons or evidences for the claim). However, some parts of an argument may be implicit or may be simply missing.

Notice that you should also classify the text as argumentative, if the speaker argues about a topic that is not related to the discussed issue (the question itself). However, you should be careful if some background information has been stated, but it is not part of the argumentation.

Notice also that not every sentence in the document has to be classified as a premise or a claim, some sentences are simply not-argumentative. Those sentences have to be marked with the label *"Non-Arg"*.

### B.2.2 *Annotation Level and Splitting Rule*

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 annotate argument components at the clause-level. For example, "We cannot make accurate predictions right now because it is still too early": here both the claim "We cannot make accurate predictions right now" and its premise "because it is still too early" are stated in one sentence.

If we have complete statements in the same sentence, we should only consider them as different argument components if there is a sign of a relation of *inference* between them. Particularly, statements connected with conjunctions like "and" or "or" usually do not include an inference step, conditional sentences as well (if, then). Instead, inference could appear in the following forms [SG14a]:

"Claim *because of* premise"
"*Given the fact* premise, *then* claim."
"*Since* premise, it *results* that claim"

However, no overlapping between the components is allowed. Similarly, no multi-label is allowed.

## B.3 ARGUMENT COMPONENTS

### *The Claim*

The claim is the central node of an argument. It is an assertion that can either appear as a conclusion where it has the character of a consequence or as an initial assertion that is supported by reasons in the subsequent statements.

### *Claim Indicators*

We state in the following a list of claim indicators that may help the annotation process. However, note that this is not a finite set, and it is not necessarily to find one of them for each single claim.

*[Therefore, thus, hence, so, as a result, as a consequence, clearly, in conclusion, implies, indicates that, it follows that, shows that, proves that, demonstrates that we may infer that, for that reason, I think, I believe, In my opinion, Our view is, and similar].*

Note that indicators should also be covered as part of the claim sentence (or clause).

*The Premise*

The premise is the arm which the arguer uses to defend his standpoint, and to increase its acceptability.

The premise should be connected to the claim topic, either directly or indirectly. Meaning that, you can find a chain of reasoning (a premise is supporting another one, till reaching the final claim) or a convergent reasoning (where each single premise is directly supporting the point of the claim). We accept any sort of argument reasoning form. Therefore, a statement is a premise if it is:

- A reason or a justification for the considered claim.
- Supporting another premise.
- Contributing to the confirmation of the claim.
- Some premises appear as an opponent statement, where the arguer discusses the opposite point of view and defeat it. In this case, this is still a premise but with an *attack* relation to the claim, instead of a *support* in the normal case.

*Premise Indicators*

Similarly to a claim, we list in the following most common premise indicators in a conversational argumentation:
*[Because, As, since, In addition, whereas, due to, furthermore, in light of, considering that, assuming that, for example, due to, given that, as shown in, and similar]*

## B.4 RELATIONS BETWEEN ARGUMENT COMPONENTS

We adopt the most frequent relation among argument components, which is :a premise to a claim. Every premise should be connected to its related claim by either a *support* or *attack* relation.

A support statement could be a justification or a reason whereas an attack statement highlights a weakness (or against) point with respect to the target claim. This attacked relation is used in rare cases as opposit to the support one. The speaker may state some attacking premises to his final claim, in order to discuss the opposed view and prevent any future criticism. That strategy may lead to a more convincing argument.

Formulate the fragment of text using «it is true that sentence 2 because sentence 1», if it's meaningful, then it's a support relation. Formulate the sentence using «it is not true that sentence 2 because sentence 1», and if it's meaningful, then it's an attack relation.

Note that our focus of study is the micro-structure of an argument. Hence, we do not require you to annotate the external relations between arguments (i.e., claim to claim). Similarly, we also do not consider the premise to premise relation. We rather consider all connected premises as a chain of support to the final claim.

# C

# GUIDELINES FOR ANNOTATING ARGUMENT QUALITY

## C.1 INTRODUCTION

Argument quality assessment is the practice of ranking the argument based on its strength and weakness traits.

We define in our data two approaches of judging the argument: the first one is based on assessing the whole argumentation as one unit, while the second approach is based on recognizing the type of each of the argument components.

You can have an overview about the task in Figure c.1

## C.2 QUALITY OF THE OVERALL ARGUMENT (PREMISES+CLAIM)

In this section, we consider the quality of the complete argument (composed by all its claim, premises, and relations).

Table c.1 details the definition and scale of assessment for five argument dimensions: Specificity, Persuasiveness, Strength, Objectivity, and Temporal-history.

For simplicity and compatibility reasons, we design those quality metrics to be added as a relation between the last premise and the argument's claim. [1] In other words, the annotator has to add those five argument quality metrics to the last premise-claim relation but considering the complete argument (i.e., all premises and relations). We will apply some examples in our training sessions.

> Please note that, in case there are more than one premise to the same claim, you need to attach the quality metrics to only the *last* premise. Yet, support/attack relations have to be assigned between every premise and the claim.

1 Adding a *list* of choices (e.g., strong-0, strong-1, strong-2) to label studio was only possible through units relations. Therefore, we design it along with the support/attack relation itself to ease the task for the annotators.

Figure c.1: Taxonomy of argument mining tasks according to *FinArgQuality*. The three tasks of: argument identification, argument unit classification, and argument relation classification can be done using *FinArg* as well.

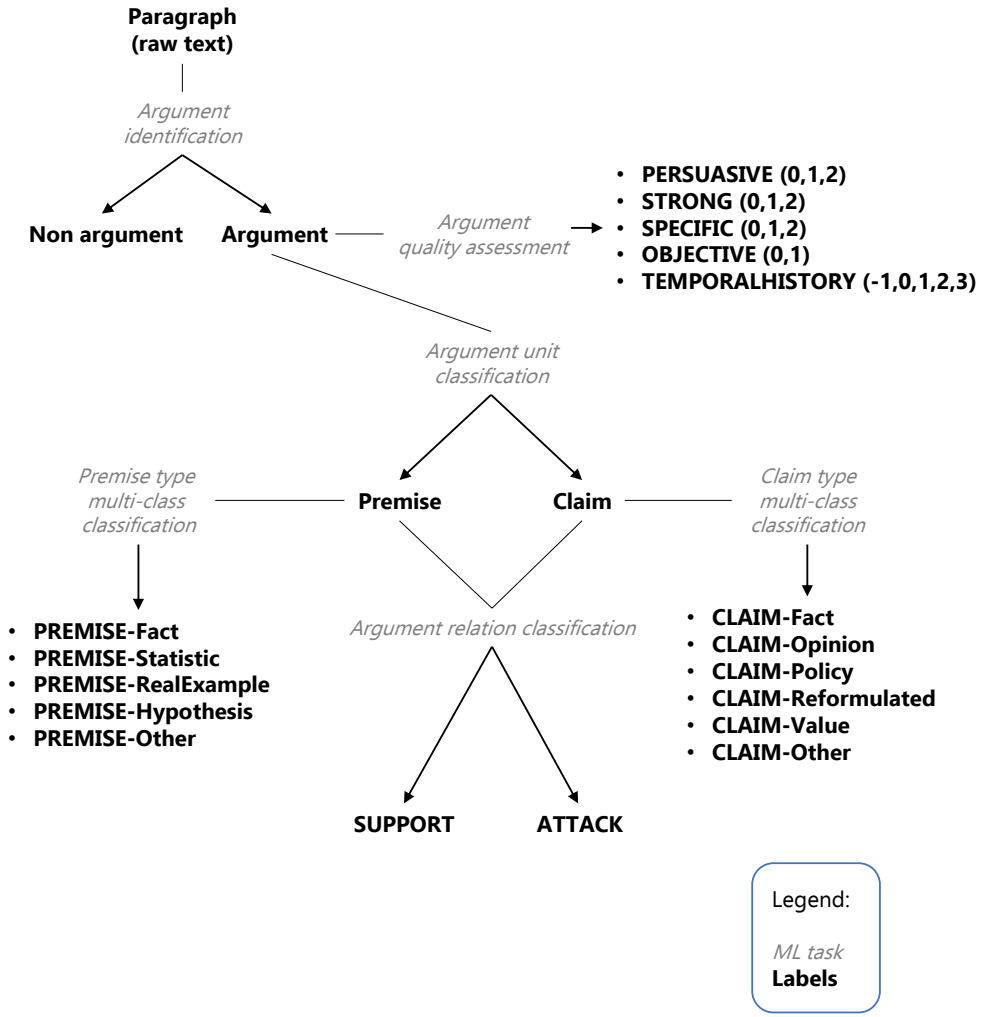


Table c.1: Quality dimensions at the argument level.

| Attribute | Definition | Score |
|---|---|---|
| **Specificity** | 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).<br>• Specific-1: The statement partially answers the question, but still implies some hedging.<br>• Specific-2: The argument is concrete and directly related to the question. |
| **Persuasiveness** | 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.<br>• Persuasive-1: The argument provides acceptable reasoning, may still contain some defects that decrease its ability of convincing. Hence, it would persuade some listeners.<br>• 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. |
| **Strength** | 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).<br>• Strong-1: A decent, fairly clear argument. The argument has at least two premises that authorize its standpoint.<br>• Strong-2: A clear and well-defended argument, supported by concrete and powerful premises. |
| **Objectivity** | Is the argument based on facts rather than feelings or opinions? | • Objective-0: A subjective or biased argument based on particular views and opinions.<br>• 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<br>• Temporal-2: up to two quarters<br>• Temporal-1: half to one year<br>• Temporal-0: more than one year<br>• Temporal-1: not mentioned (If there is no explicit time indicator choose this value, even if you feel that it could be concluded from the context). |

In addition, you can see in Table c.2 some examples of temporal_history annotation:

Table c.2: Examples on the temporal-history dimension of an argument

| Score | Description | Example |
|---|---|---|
| 3 | recent (during this quarter) | "It's important to note that it's not just large brand advertisers that are doing video, but all of our market segments. Direct response, SMBs who have uploaded 1.5 million videos and have both organic and paid in the *last month*, and developers." (2015-Q4) |
| 2 | short-past (up to 2 quarters) | "Yes, a good example of the first-line opportunity was something that you could have seen at NRF this *January*. We launched, for example, Teams for first-line workers, which had things like shift worker capabilities, secure messaging." (2019-Q2) |
| 1 | mid-past (half to one year) | "So it will fluctuate quarter-to-quarter. I would say last year in the first half was a pretty large investment area. I'll lump it in with capital expenditures, but in the first two quarters, Q1 of last year, it was 82% growth year-over-year in capital expenditures, Q2 was 67%. This year, those numbers are 33% in Q1, and 1% in Q2" (2018-Q2) |
| 0 | long-past (more than 1 year) | "Of course, we're also improving the quality of the services that we provide and if you look back during the last three years, we've added new services to our portfolio. We added Apple Pay, we added Apple Music. We added this advertising business on our App Store." (2018-Q4) |
| -1 | Not mentioned | "So, when you think about segments, the majority of these relate to the AWS segment. And it's tech infrastructure assets. So, it's – I should say the servers are tech infrastructure assets."(2019-Q4) |

## C.3 QUALITY OF ARGUMENT COMPONENTS

### C.3.1 *Premise Types*

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

   > 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."*

2. **Real example**: A past event or a similar, related example (e.g., another product status).

   > 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."*

3. **Statistics**: technical indicators or any statistical numbers.

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

4. **Hypothesis**: an assumption made to draw out its consequences.

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

5. **Other**: if no one of the previous types matches.

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

### C.3.2 *Claim Types*

The claim type is the category of what is being claimed. According to our observations, and inline with previous works [Car+18], we define the following possible types of a claim:

1. **Fact**: something is or is not the case.

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

2. **Value**: discuss a quantities measure like returns, sales, or similar.

> 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."*

3. **Policy**: argues that certain conditions should exist, or that something should or should not be taken, in order to solve a problem or change the course of the action.

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

4. **Opinion**: A feeling, a belief, or way of thinking about something. 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*.

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

5. **Reformulated**: This is a label to point to the same previously mentioned claim. This comes to the scene frequently in oral argumentation where the speaker states his claims, the premises for it, then re-insist on the

result by rewording the same claim. In this case, we annotate this claim as a "Reformulated" without the need to link it to the past premises by any relation (We link them only to the original claim). In other words, we annotate the reformulated claim in our data as a single claim with *no premises*. The reformulated claim is mostly the shorter one of the two claims. Few terms introducing reformulations are: in other words, that is to say, rather, and it can appear without any preamble.

> 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."*

6. **Other**: if none of the previous types has matched.

The annotation is to be done using Label Studio (see Appendix d).

# d

## LABEL STUDIO

We use for our annotation the free tool of Label Studio. Before we upload the data, we had set up and prepared the data with the following steps:

- For each transcript, we created a list of all the speakers. We call those Pre-assigned labels (which the annotators still can correct in case of wrong automatic tokenization):
  - Operator
  - Analyst
  - Representative
  - Intro
  - Question
  - Answer
- We define a set of labels for the annotators to mark each unit with, according to our guidelines: premise types, claim types and Non-argument (see Figure d.1).
- Additionally, we define the possible relations between the premise and the claim: Support or Attack. As well as the relation representing the argument quality criteria.

The output file generated by Label Studio is in a JSON format (see Figure d.2), from which we extract the main information to be in an annotation file (see Figure d.3).

Figure d.1: An example (screenshot) of Label Studio API covering the argument structure and the argument quality dimensions.

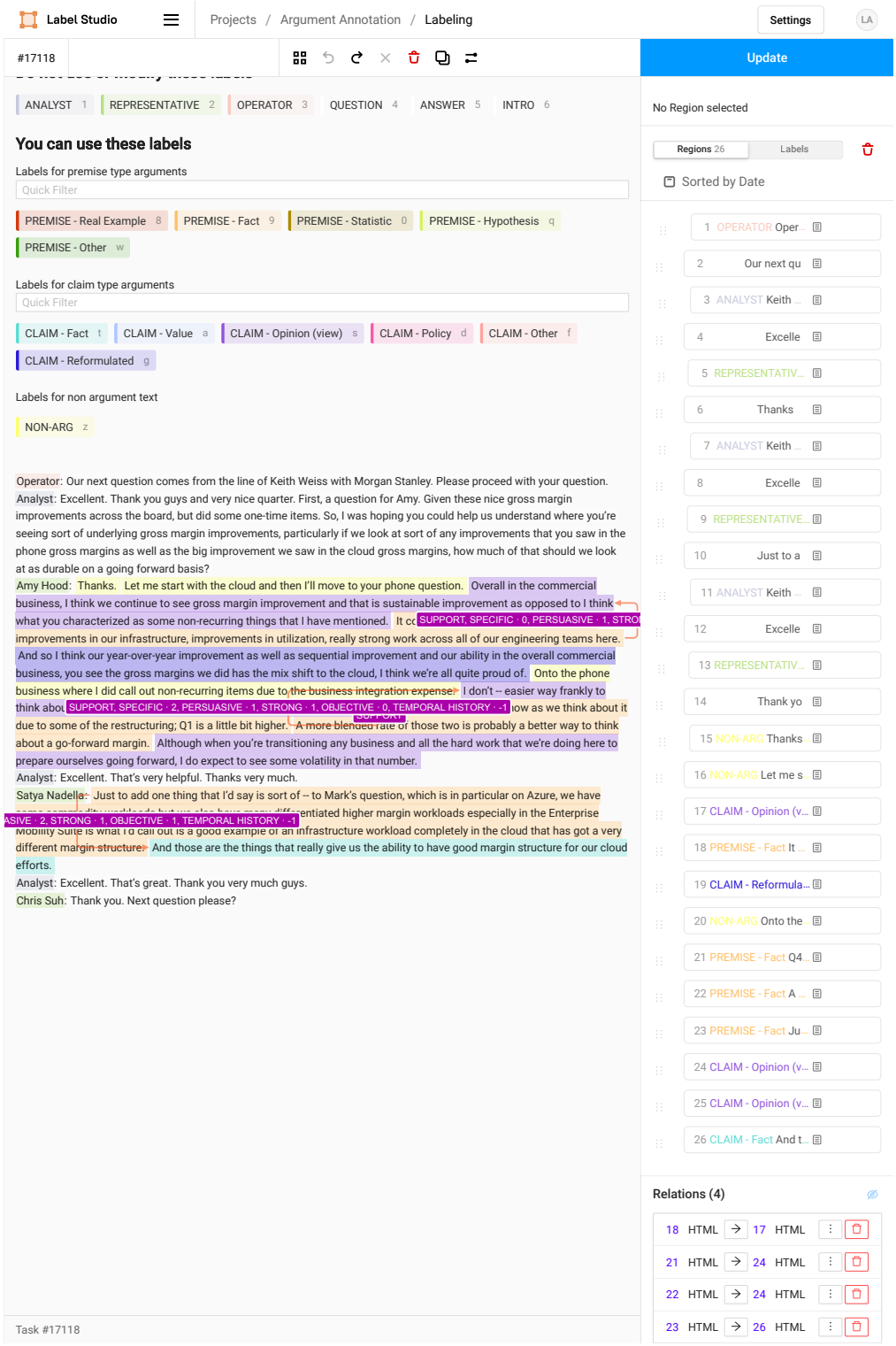


Figure d.2: Part of the JSON file corresponding to the example d.1.

```
{
  "id": 17118,
  "data": {
    "year": 2015,
    "company": "MSFT",
    "my_text": ...,
    "quarter": 1
  },
  "annotations": [
    {
      "id": 20926,
     ...
      "result": [
        {
          "id": "GlKaImx9rY",
          "type": "labels",
          "value": {
            "end": 9,
            "text": "Operator",
            "start": 0,
            "labels": [
              "OPERATOR"
            ]
          },
          "origin": "manual"
        },
        {
          "id": "DWXt7xqwX6",
          "type": "labels",
          "value": {
            "end": 119,
            "text": " Our next question comes from the line of Keith Weiss with Morgan
                 Stanley. Please proceed with your question.",
            "start": 10,
            "labels": [
              "INTRO"
            ]
        ,
        %ANALYST
        %REPRESENTATIVE
           ......
            {
          "id": "kWdB7Uf39p",
          "type": "labels",
          "value": {
            "end": 735,
            "text": "Let me start with the cloud and then I'll move to your phone question.",
            "start": 665,
            "labels": [
              "NON-ARG"
          ....
      }
         ]
          },
       {
          "id": "qaBS6CYG1z",
          "type": "labels",
          "value": {
            "end": 956,
            "text": "Overall in the commercial business, I think we continue to see gross
                 margin improvement...",
            "start": 736,
            "labels": [
              "CLAIM - Opinion (view)"
            ]
          },
           ....}]
      ] }
```

Figure d.3: Part of the generated annotation file corresponding to the same example in Figures d.1 and d.2.

- L1GB7MuZfL **NON-ARG** 651 664 Thanks.
- kWdB7Uf39p **NON-ARG** 665 735 Let me start with the cloud and then I'll move to your phone question.
- qaBS6CYG1z **CLAIM-Opinion(view)** 736 956 Overall in the commercial business, I think we continue to see gross margin improvement .. I have mentioned.
- NFHaBlmVSM **PREMISE-Fact** 957 1125 It continues to be improvements in scale, improvements in our infrastructure, ... engineering teams here.
- AAAZ9y8SNu **CLAIM-Reformulated** 1125 1355 And so I think our year-over-year improvement ... I think we're all quite proud of.
- ZwGgVxXflK **NON-ARG** 1356 1461 Onto the phone business where I did call out non-recurring items due to the business integration expense.
- PbWhy7VVgM **PREMISE-Fact** 1566 1680 Q4 was a little depressed is how as we think about it due to some of the restructuring; Q1 is a little bit higher.
- .....
- cHeGdAIofx **NON-ARG** 2615 2636 Next question please?
- **R0** SUPPORT ARG1:NFHaBlmVSM ARG2:qaBS6CYG1z
- **R1** SUPPORT ARG1:PbWhy7VVgM ARG2:nft9yZyTzo
- **R2** SUPPORT ARG1:SlZq2ec8uY ARG2:nft9yZyTzo
- **R3** SUPPORT ARG1:FhJhDmj7y3 ARG2:qa2VwpVFaa
- **ArgQ0** qaBS6CYG1z SPECIFIC_0 PERSUASIVE_1 STRONG_1 OBJECTIVE_1 TEMPORALHISTORY_1
- **ArgQ1** nft9yZyTzo SPECIFIC_2 PERSUASIVE_1 STRONG_1 OBJECTIVE_0 TEMPORALHISTORY_-1
- **ArgQ2** qa2VwpVFaa SPECIFIC_1 PERSUASIVE_2 STRONG_1 OBJECTIVE_1 TEMPORALHISTORY_-1

# BIBLIOGRAPHY


[AST20] Tinsaye Abye, Tilmann Sager, and Anna Juliane Triebel. "An Open-Domain Web Search Engine for Answering Comparative Questions." In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_130.pdf (cit. on p. 101).

[Aga20] Arul Agarwal. "Sentiment Analysis of Financial News." In: *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE. 2020, pp. 312–315 (cit. on p. 73).

[Aha+14] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. "A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics." In: *Proceedings of the first workshop on argumentation mining*. 2014, pp. 64–68 (cit. on p. 23).

[Ajj+17] Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. "Unit segmentation of argumentative texts." In: *Proceedings of the 4th Workshop on Argument Mining*. 2017, pp. 118–128 (cit. on pp. 38, 49).

[Ajj+18] Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. "Visualization of the Topic Space of Argument Search Results in args.me." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 60–65. DOI: 10.18653/v1/D18-2011. URL: https://aclanthology.org/D18-2011 (cit. on p. 101).

[Ajj+19] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. "Data acquisition for argument search: The args. me corpus." In: *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer. 2019, pp. 48–59 (cit. on p. 101).

[AK19] Khalid Al-Khatib. "Computational Analysis of Argumentation Strategies." Dissertation. Bauhaus-Universität Weimar, Dec. 2019 (cit. on pp. 16, 79).

[AK+16a] Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. "A news editorial corpus for mining argumentation strategies." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 3433–3443 (cit. on pp. 32, 38, 78).

[AK+16b] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, Benno Stein, and Steve Göring. *Webis-Editorials-16*. Dec. 2016. DOI: 10.5281/zenodo.3254405. URL: https://doi.org/10.5281/zenodo.3254405 (cit. on p. 112).

[Alh+21a] Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, and Jelena Mitrović. "DistilBERT-based Argumentation Retrieval for Answering Comparative Questions." In: *Working Notes of CLEF* (2021) (cit. on pp. 103, 107).

[Alh+21b] Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. "A Stacking Approach for Cross-Domain Argument Identification." In: *Database and Expert Systems Applications*. Ed. by Christine Strauss, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil. Cham: Springer International Publishing, 2021, pp. 361–373. ISBN: 978-3-030-86472-9 (cit. on pp. 54, 68, 100).

[Alh+22a] Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. "Query Expansion, Argument Mining and Document Scoring for an Efficient Question Answering System." In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2022, pp. 162–174 (cit. on pp. 88, 107).

[Alh+22b] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. "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): Association for Computational Linguistics, Dec. 2022, pp. 163–169. URL: https://aclanthology.org/2022.finnlp-1.22 (cit. on p. 60).

[ALEZ22] 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*. 2022 (cit. on p. 35).

[Alh+21c] 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*. Springer. 2021, pp. 15–26 (cit. on pp. 32, 34, 73).

[All16] Jens Allwood. "Argumentation, Activity and Culture." In: *COMMA*. 2016, p. 3 (cit. on p. 72).

[AMP00] Leila Amgoud, Nicolas Maudet, and Simon Parsons. "Modelling dialogues using argumentation." In: *Proceedings Fourth International Conference on MultiAgent Systems*. IEEE. 2000, pp. 31–38 (cit. on p. 20).

[AP08] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics." In: *Computational linguistics* 34.4 (2008), pp. 555–596 (cit. on p. 66).

[AD19] Hiteshwar Kumar Azad and Akshay Deepak. "Query expansion techniques for information retrieval: a survey." In: *Information Processing & Management* 56.5 (2019), pp. 1698–1735 (cit. on p. 101).

[BX22] Sudipta Basu and Zhongnan Xiang. "Why Don't Analysts Always Value Earnings Conference Calls?" In: *Available at SSRN* (2022) (cit. on p. 96).

[Bea50] Monroe C Beardsley. *Practical Logic. New Yorks Prentice-Hall*. 1950 (cit. on p. 61).

[BMB10] Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. "A taxonomy of argumentation models used for knowledge representation." In: *Artificial Intelligence Review* 33.3 (2010), pp. 211–259 (cit. on pp. 15–17, 20, 76).

[Bev+18] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. "Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl." In: *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*. Ed. by Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer, Mar. 2018 (cit. on p. 103).

[BR11] Or Biran and Owen Rambow. "Identifying justifications in written dialogs by classifying text as argumentative." In: *International Journal of Semantic Computing* 5.04 (2011), pp. 363–381 (cit. on p. 45).

[Bon+22] Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. "Towards understanding and answering comparative questions." In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 2022, pp. 66–74 (cit. on p. 99).

[Bon+20a] Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. "Comparative web search questions." In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 2020, pp. 52–60 (cit. on p. 99).

[Bon+21a] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Alexander Panchenko Yamen Ajjour, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. "Overview of Touché 2021: Argument Retrieval." In: *Advances in Information Retrieval. 43rd European Conference on IR Research (ECIR 2021)*. Ed. by Djoerd Hiemstra, Maria-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani. Vol. 12036. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer, Mar. 2021, pp. 574–582. DOI: 10.1007/978-3-030-72240-1_67. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1_67 (cit. on p. 101).

[Bon+20b] Alexander Bondarenko, Alexander Panchenko, Meriem Beloucif, Chris Biemann, and Matthias Hagen. "Answering Comparative Questions with Arguments." In: *Datenbank-Spektrum* 20.2 (2020), pp. 155–160. DOI: 10.1007/s13222-020-00346-8 (cit. on p. 100).

[Bon+21b] Alexander Bondarenko et al. "Overview of Touché 2021: Argument Retrieval." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021)*. Ed. by K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro. Vol. 12880. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer, Sept. 2021, pp. 450–467. DOI: 10.1007/978-3-030-85251-1_28. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_28 (cit. on pp. 99, 100, 107).

[BAA19] Rihab Bouslama, Raouia Ayachi, and Nahla Ben Amor. "Using Convolutional Neural Network in Cross-Domain Argumentation Mining Framework." In: *International Conference on*

*Scalable Uncertainty Management*. Springer. 2019, pp. 355–367 (cit. on p. 39).

[Bro+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 110).

[BPQ20] Oscar Bustos and A Pomares-Quimbaya. "Stock market movement forecast: A Systematic review." In: *Expert Systems with Applications* 156 (2020), p. 113464 (cit. on pp. 33, 34).

[Car+18] Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. "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: Association for Computational Linguistics, July 2018, pp. 621–631. DOI: 10.18653/v1/P18-1058. URL: https://aclanthology.org/P18-1058 (cit. on pp. 74, 75, 77, 78, 127).

[Cas+20] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. "I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language." In: *Proceedings of LREC*. 2020 (cit. on p. 43).

[CP21] Viktoriia Chekalina and Alexander Panchenko. "Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval." In: *Working Notes of CLEF* (2021) (cit. on pp. 99–101).

[CHC18] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "NTUSD-Fin: a market sentiment dictionary for financial social media data applications." In: *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*. 2018, pp. 37–43 (cit. on p. 32).

[CHC20a] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "Issues and perspectives from 10,000 annotated financial social media data." In: *Proceedings of The 12th language resources and evaluation conference*. 2020, pp. 6106–6110 (cit. on p. 32).

[CHC20b] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "NLP in FinTech applications: past, present and future." In: *arXiv preprint arXiv:2005.01320* (2020) (cit. on pp. 26, 63).

[CHC20c] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "NumClaim: Investor's Fine-grained Claim Detection." In: *Proceedings of the 29th ACM International Conference on Informa-*

*tion & Knowledge Management*. 2020, pp. 1973–1976 (cit. on p. 36).

[CHC21a] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "Evaluating the rationales of amateur investors." In: *Proceedings of the Web Conference 2021*. 2021, pp. 3987–3998 (cit. on pp. 31, 60, 73, 76).

[CHC21b] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From Opinion Mining to Financial Argument Mining*. Springer Nature, 2021 (cit. on pp. 8, 9, 27, 28, 30, 35).

[Che+22] 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 (cit. on p. 35).

[Che+18] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. "Numeral understanding in financial tweets for fine-grained crowd-based forecasting." In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2018, pp. 136–143 (cit. on pp. 30, 76, 79).

[Che+19] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. "Targer: Neural argument mining at your fingertips." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 195–200 (cit. on pp. 99, 101).

[Chi20] Nathan Chiu. "The Impact of Individual and Collective Attribution on Earnings Calls Impression Management By." In: (2020) (cit. on pp. 28, 59).

[Cho+22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. "Palm: Scaling language modeling with pathways." In: *arXiv preprint arXiv:2204.02311* (2022) (cit. on p. 110).

[Cic+18] Giovanni Ciccone, Arthur Sultan, Léa Laporte, Elod Egyed-Zsigmond, Alaa Alhamzeh, and Michael Granitzer. "Stacked gender prediction from tweet texts and images notebook for pan at CLEF 2018." In: *CLEF 2018-Conference and Labs of the Evaluation*. 2018, 11p (cit. on p. 39).

[Coh60] Jacob Cohen. "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46 (cit. on p. 80).

[CMN20] Lauren Cohen, Christopher Malloy, and Quoc Nguyen. "Lazy prices." In: *The Journal of Finance* 75.3 (2020), pp. 1371–1415 (cit. on p. 27).

[Coh87] Robin Cohen. "Analyzing the structure of argumentative discourse." In: *Computational linguistics* 13 (1987), pp. 11–24 (cit. on pp. 61, 111).

[CH+21] Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. "Combining pre-trained language models and structured knowledge." In: *arXiv preprint arXiv:2101.12294* (2021) (cit. on p. 109).

[CCR16] Irving Copi, Carl Cohen, and Victor Rodych. *Introduction to logic*. Routledge, 2016 (cit. on p. 24).

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. on p. 10).

[CTW06] Julie Cotter, Irem Tuna, and Peter D Wysocki. "Expectations management and beatable targets: How do analysts react to explicit earnings guidance?" In: *Contemporary accounting research* 23.3 (2006), pp. 593–624 (cit. on p. 28).

[CC18] Belinda Crawford Camiciottoli. "Persuasion in earnings calls: A diachronic pragmalinguistic analysis." In: *International Journal of Business Communication* 55.3 (2018), pp. 275–292 (cit. on pp. 76, 77, 81).

[Dam12] T Edward Damer. *Attacking faulty reasoning*. Cengage Learning, 2012 (cit. on pp. 24, 61).

[Dax+20] Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. "Argumentext: argument classification and clustering in a generalized search scenario." In: *Datenbank-Spektrum* 20.2 (2020), pp. 115–121 (cit. on p. 100).

[DW13] Joost CF De Winter. "Using the Student's t-test with extremely small sample sizes." In: *Practical Assessment, Research, and Evaluation* 18.1 (2013), p. 10 (cit. on p. 48).

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on pp. 38, 42, 43, 85, 88, 110).

[Din+20] Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. "Cogltx: Applying bert to long texts." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12792–12804 (cit. on p. 86).

[Din+15] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. “Deep learning for event-driven stock prediction.” In: *Twenty-fourth international joint conference on artificial intelligence*. 2015 (cit. on p. 34).

[DLC20] Esin Durmus, Faisal Ladhak, and Claire Cardie. “The role of pragmatic and discourse context in determining argument impact.” In: *arXiv preprint arXiv:2004.03034* (2020) (cit. on p. 86).

[EH02] Frans H van Eemeren and Peter Houtlosser. “Strategic maneuvering.” In: *dialectic and rhetoric*. Springer, 2002, pp. 131–159 (cit. on pp. 9, 71).

[EM15] Matthias Eickhoff and Jan Muntermann. “Stock analysts vs. the crowd: a study on mutual prediction.” In: (2015) (cit. on p. 31).

[EHV21] Aparna Elangovan, Jiayuan He, and Karin Verspoor. “Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance Evaluation.” In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1325–1335. DOI: 10.18653/v1/2021.eacl-main.113. URL: https://aclanthology.org/2021.eacl-main.113 (cit. on p. 39).

[Est+10] Fernando Estrada et al. “Theory of argumentation in financial markets.” In: *Journal of Advanced Studies in Finance (JASF)* 1.01 (2010), pp. 18–22 (cit. on p. 35).

[Fam95] Eugene F Fama. “Random walks in stock market prices.” In: *Financial analysts journal* 51.1 (1995), pp. 75–80 (cit. on p. 34).

[FSB15] Noura Farra, Swapna Somasundaran, and Jill Burstein. “Scoring persuasive essays using opinions and their targets.” In: *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*. 2015, pp. 64–74 (cit. on p. 72).

[Fis+22] Irina Fishcheva, Dmitriy Osadchiy, Klavdiya Bochenina, and Evgeny Kotelnikov. “Argumentative Text Generation in Economic Domain.” In: *arXiv preprint arXiv:2206.09251* (2022) (cit. on pp. 8, 35, 98).

[FLP13] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013 (cit. on p. 80).

[Fre11] James B. Freeman. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter Mouton, 2011. ISBN: 9783110875843. DOI: doi:10.1515/9783110875843. URL: https://doi.org/10.1515/9783110875843 (cit. on pp. 21, 60).

[FJ90] Noah E Friedkin and Eugene C Johnsen. "Social influence and opinions." In: *Journal of Mathematical Sociology* 15.3-4 (1990), pp. 193–206 (cit. on p. 96).

[Fu+21] Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan, and Simon Corston-Oliver. "Improving Punctuation Restoration for Speech Transcripts via External Data." In: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Online: Association for Computational Linguistics, Nov. 2021, pp. 168–174. DOI: 10.18653/v1/2021.wnut-1.19. URL: https://aclanthology.org/2021.wnut-1.19 (cit. on p. 65).

[Gou+19] Régis Goubin, Dorian Lefeuvre, Alaa Alhamzeh, Jelena Mitrovic, Elöd Egyed-Zsigmond, and Leopold Ghemmogne Fossi. "Bots and Gender Profiling using a Multi-layer Architecture." In: *CLEF (Working Notes)*. 2019 (cit. on p. 39).

[Gov01] Trudy Govier. "A Practical Study of Argument, (Belmont." In: *CA: Wadsworth* (2001) (cit. on p. 14).

[Gov10] Trudy Govier. *A practical study of argument*. Cengage Learning, 7th edition, 2010 (cit. on p. 61).

[Gre+20] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. "A large-scale dataset for argument quality ranking: Construction and analysis." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7805–7813 (cit. on pp. 86, 93).

[Gro96] Leo Groarke. "Informal logic." In: (1996) (cit. on p. 24).

[GVE04] Rob Grootendorst and Frans H Van Eemeren. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004 (cit. on pp. 15, 25).

[GLC22] Chong Guan, Wenting Liu, and Jack Yu-Chao Cheng. "Using social media to predict the stock market crash and rebound amid the pandemic: the digital 'haves' and 'have-mores'." In: *Annals of Data Science* 9.1 (2022), pp. 5–31 (cit. on p. 31).

[Gup+20] Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. "Comprehensive review of text-mining applications in finance." In: *Financial Innovation* 6.1 (2020), pp. 1–25 (cit. on p. 26).

[GAW21] Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. "Assessing the Sufficiency of Arguments through Conclusion Generation." In: *arXiv preprint arXiv:2110.13495* (2021) (cit. on p. 72).

[HG16] Ivan Habernal and Iryna Gurevych. “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*. 2016, pp. 1214–1223 (cit. on pp. 25, 72).

[HG17] Ivan Habernal and Iryna Gurevych. “Argumentation mining in user-generated web discourse.” In: *Computational Linguistics* 43.1 (2017), pp. 125–179 (cit. on pp. 9, 19–22, 38, 39, 60, 61, 63, 68, 104).

[Hab+18] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. “Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation.” In: *arXiv preprint arXiv:1802.06613* (2018) (cit. on p. 24).

[Has62] Arthur Claude Hastings. *A Reformulation of the Modes of Reasoning in Argumentation*. Northwestern University, 1962 (cit. on p. 17).

[Hen00] A Henkemans. “State-of-the-art: The structure of argumentation.” In: *Argumentation* 14.4 (2000), pp. 447–473 (cit. on pp. 19, 20, 66, 80).

[Hit03] David Hitchcock. “Toulmin’s warrants.” In: *Anyone who has a view*. Springer, 2003, pp. 69–82 (cit. on p. 21).

[Hu+18] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. “Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction.” In: *Proceedings of the eleventh ACM international conference on web search and data mining*. 2018, pp. 261–269 (cit. on p. 34).

[Huc20] Johannes Huck. “Development of a Search Engine to Answer Comparative Queries.” In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: `http://ceur-ws.org/Vol-2696/paper\_178.pdf` (cit. on p. 101).

[Hur11] Kristian Hursti. “Management earnings forecasts: Could an investor reliably detect an unduly positive bias on the basis of the strength of the argumentation?” In: *The Journal of Business Communication (1973)* 48.4 (2011), pp. 393–408 (cit. on pp. 8, 35).

[Hut05] Amy P Hutton. “Determinants of managerial earnings guidance prior to regulation fair disclosure and bias in analysts’ earnings forecasts.” In: *Contemporary Accounting Research* 22.4 (2005), pp. 867–914 (cit. on p. 28).

[JB97] Ralph Henry Johnson and J Anthony Blair. *Logical self-defense*. McGraw-Hill Ryerson, 1997 (cit. on p. 24).

[JB06] Ralph Henry Johnson and J Anthony Blair. *Logical self-defense*. International Debate Education Association, 2006 (cit. on pp. 9, 24, 74).

[JSAH07] Shafiq Rayhan Joty and Sheikh Sadid-Al-Hasan. "Advances in focused retrieval: A general review." In: *2007 10th international conference on computer and information technology*. IEEE. 2007, pp. 1–5 (cit. on p. 100).

[KL14] Colm Kearney and Sha Liu. "Textual sentiment in finance: A survey of methods and models." In: *International Review of Financial Analysis* 33 (2014), pp. 171–185 (cit. on p. 32).

[KS19] Katherine A Keith and Amanda Stent. "Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls." In: *arXiv preprint arXiv:1906.02868* (2019) (cit. on pp. 8, 11, 28, 73, 74, 93, 94, 98).

[Ken+06] George A Kennedy et al. "On rhetoric: A theory of civic discourse." In: (2006) (cit. on p. 24).

[KH06] Soo-Min Kim and Eduard Hovy. "Extracting opinions, opinion holders, and topics expressed in online news media text." In: *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. 2006, pp. 1–8 (cit. on p. 32).

[KD97] Alistair Knott and Robert Dale. "Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations." In: (Aug. 1997) (cit. on p. 41).

[Kri80] Klaus Krippendorff. "Content Analysis: An Introduction to its Methodology." In: *Sage* (1980) (cit. on pp. 22, 65).

[Kri04] Klaus Krippendorff. "Measuring the reliability of qualitative text analysis data." In: *Quality and quantity* 38 (2004), pp. 787–800 (cit. on pp. 22, 65).

[Kri18] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018 (cit. on p. 80).

[LPH07] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner." In: (2007) (cit. on p. 39).

[LK77] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data." In: *biometrics* (1977), pp. 159–174 (cit. on pp. 23, 80).

[Lau21] Anne Lauscher. "Language representations for computational argumentation." In: (2021) (cit. on p. 93).

[Lau+20] Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. "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): International Committee on Computational Linguistics, Dec. 2020, pp. 4563–4574. DOI: 10.18653/v1/2020.coling-main.402. URL: https://aclanthology.org/2020.coling-main.402 (cit. on pp. 71, 72).

[Li+20] Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction." In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 3063–3070 (cit. on pp. 8, 28).

[LP20] Davide Liga and Monica Palmirani. "Transfer Learning with Sentence Embeddings for Argumentative Evidence Classification." In: (2020) (cit. on p. 38).

[LKB20] Qi Liu, Matt J Kusner, and Phil Blunsom. "A survey on contextual embeddings." In: *arXiv preprint arXiv:2003.07278* (2020) (cit. on p. 88).

[LLS09] Ying Liu, Han Tong Loh, and Aixin Sun. "Imbalanced text classification: A term weighting approach." In: *Expert systems with Applications* 36.1 (2009), pp. 690–701 (cit. on p. 90).

[Liu+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on pp. 38, 86).

[Lop+20] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. "Transformer-based end-to-end question generation." In: *arXiv preprint arXiv:2005.01107* 4 (2020) (cit. on p. 87).

[LH19] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization." In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7 (cit. on p. 43).

[LM11] Tim Loughran and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." In: *The Journal of finance* 66.1 (2011), pp. 35–65 (cit. on p. 32).

[Lum16] Christoph Lumer. "Walton's argumentation schemes." In: (2016) (cit. on p. 17).

[Mac81] Jim D Mackenzie. "The dialectics of logic." In: *Logique et analyse* 24.94 (1981), pp. 159–177 (cit. on p. 20).

[MO13] Didier Maillat and Steve Oswald. "Biases and constraints in communication: Argumentation, persuasion and manipulation." In: *Journal of pragmatics* 59 (2013) (cit. on p. 14).

[MPR11] Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. "What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions." In: *The Accounting Review* 86.4 (2011), pp. 1383–1414 (cit. on pp. 7, 28, 59).

[MS16] Khalid Al-Khatib Henning Wachsmuth Matthias and Hagen Jonas Köhler Benno Stein. "Cross-Domain Mining of Argumentative Text through Distant Supervision." In: *Proceedings of NAACL-HLT*. 2016, pp. 1395–1404 (cit. on pp. 9, 37, 38, 53, 117).

[May+20] Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. "Generating Adversarial Examples for Topic-Dependent Argument Classification 1." In: *Computational Models of Argument*. IOS Press, 2020, pp. 33–44 (cit. on p. 38).

[Maz15] Mariusz Maziarz. "A review of the Granger-causality fallacy." In: *The journal of philosophical economics: Reflections on economic and social issues* 8.2 (2015), pp. 86–105 (cit. on p. 31).

[MML20] R. Thomas McCoy, Junghyun Min, and Tal Linzen. "BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance." In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 217–227. DOI: 10.18653/v1/2020.blackboxnlp-1.21. URL: https://aclanthology.org/2020.blackboxnlp-1.21 (cit. on pp. 38, 39, 56, 110).

[Mis95] Connie A Missimer. *Good arguments: An introduction to critical thinking*. Prentice Hall, 1995 (cit. on p. 14).

[MM11] Raquel Mochales and Marie-Francine Moens. "Argumentation mining." In: *Artificial Intelligence and Law* 19.1 (2011), pp. 1–22 (cit. on p. 67).

[Moe18] Marie-Francine Moens. "Argumentation mining: How can a machine acquire common sense and world knowledge?" In: *Argument & Computation* 9.1 (2018), pp. 1–14 (cit. on p. 93).

[Moe+07] Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. "Automatic detection of arguments in legal texts." In: *Proceedings of the 11th international conference on Artificial intelligence and law*. 2007, pp. 225–230 (cit. on pp. 40, 41).

[Nag+15] Kenji Nagata, Jun Kitazono, Shinichi Nakajima, Satoshi Eifuku, Ryoi Tamura, and Masato Okada. "An exhaustive search and stability of sparse estimation for feature selection problem." In: *IPSJ Online Transactions* 8 (2015), pp. 25–32 (cit. on p. 88).

[Nas+14] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. "Text mining for market prediction: A systematic review." In: *Expert Systems with Applications* 41.16 (2014), pp. 7653–7670 (cit. on p. 33).

[NB07] Todd Neideen and Karen Brasel. "Understanding statistical tests." In: *Journal of surgical education* 64.2 (2007), pp. 93–96 (cit. on p. 90).

[NK19] Timothy Niven and Hung-Yu Kao. "Probing neural network comprehension of natural language arguments." In: *arXiv preprint arXiv:1907.07355* (2019) (cit. on p. 43).

[OCA13] Nuno Oliveira, Paulo Cortez, and Nelson Areal. "On the predictability of stock market behavior using stocktwits sentiment and posting volume." In: *Portuguese conference on artificial intelligence*. Springer. 2013, pp. 355–365 (cit. on p. 31).

[PM09] 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*. 2009, pp. 98–107 (cit. on p. 60).

[Pal17] Rudi Palmieri. "The role of argumentation in financial communication and investor relations." In: *Handbook of financial communication and investor relations* (2017), pp. 45–60 (cit. on pp. 8, 35).

[PY09] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning." In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359 (cit. on p. 42).

[Pan+18] Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. "Categorizing comparative sentences." In: *arXiv preprint arXiv:1809.06152* (2018) (cit. on p. 100).

[Paz+19] 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." In: *Intelligenza Artificiale* 13.2 (2019), pp. 173–188 (cit. on pp. 8, 9, 35).

[PS13] Andreas Peldszus and Manfred Stede. "From argument diagrams to argumentation mining in texts: A survey." In: *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7.1 (2013), pp. 1–31 (cit. on p. 60).

[PN15] Isaac Persing and Vincent Ng. "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)*. 2015, pp. 543–552 (cit. on pp. 72, 74, 84, 85).

[Phi16] Thomas Philippon. *The fintech opportunity*. Tech. rep. National Bureau of Economic Research, 2016 (cit. on p. 7).

[PVN11] Barbara Plank and Gertjan Van Noord. "Effective measures of domain similarity for parsing." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011, pp. 1566–1576 (cit. on p. 37).

[Pot+19] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. "TIRA Integrated Research Architecture." In: *Information Retrieval Evaluation in a Changing World*. Ed. by Nicola Ferro and Carol Peters. The Information Retrieval Series. Berlin Heidelberg New York: Springer, Sept. 2019. ISBN: 978-3-030-22948-1. DOI: 10.1007/978-3-030-22948-1_5 (cit. on p. 101).

[Pot+12] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. "ChatNoir: A Search Engine for the ClueWeb09 Corpus." In: *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2012)*. Ed. by Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson. ACM, Aug. 2012, p. 1004. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348429 (cit. on p. 103).

[Pri+12] 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." In: *Journal of Banking & Finance* 36.4 (2012), pp. 992–1011 (cit. on pp. 7, 28, 59).

[PH14] Anna Prokofieva and Julia Hirschberg. "Hedging and speaker commitment." In: *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*. 2014 (cit. on pp. 74, 75).

[QY19] Yu Qin and Yi Yang. "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*. 2019, pp. 390–401 (cit. on pp. 28, 73).

[RR05] Chris Reed and Glenn Rowe. "Translating Toulmin diagrams: Theory neutrality in argument representation." In: *Argumentation* 19.3 (2005), pp. 267–286 (cit. on p. 20).

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084 (cit. on p. 105).

[Rei+19] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. "Classification and clustering of arguments with contextualized word embeddings." In: *arXiv preprint arXiv:1906.09821* (2019) (cit. on pp. 43, 86).

[Rin+15] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. "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: Association for Computational Linguistics, Sept. 2015, pp. 440–450. DOI: 10.18653/v1/D15-1050. URL: https://aclanthology.org/D15-1050 (cit. on p. 72).

[RZT04] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. "Simple BM25 extension to multiple weighted fields." In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. 2004, pp. 42–49 (cit. on p. 103).

[RRP19] Andrea Rocci, Carlo Raimondo, and Daniele Puccinelli. "Evidentiality and Disagreement in Earnings Conference Calls: Preliminary Empirical Findings." In: 2019, pp. 100–104 (cit. on p. 73).

[RKR20] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A primer in bertology: What we know about how bert works." In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 842–866 (cit. on p. 58).

[SR18] Omer Sagi and Lior Rokach. "Ensemble learning: A survey." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1249 (cit. on p. 39).

[SC06] Victor Sampson and Douglas Clark. "Assessment of argument in science education: A critical review of the literature." In: (2006) (cit. on p. 22).

[San+19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *arXiv preprint arXiv:1910.01108* (2019) (cit. on pp. 10, 38, 43, 100).

[Sat+15] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. "End-to-end argument generation system in debating." In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. 2015, pp. 109–114 (cit. on p. 16).

[Sau94] Kurt M Saunders. "Law as rhetoric, rhetoric as argument." In: *Journal of Legal Education* 44.4 (1994), pp. 566–578 (cit. on pp. 9, 15, 71).

[Sch+19] Matthias Schildwächter, Alexander Bondarenko, Julian Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. "Answering Comparative Questions: Better than Ten-Blue-Links?" In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 2019, pp. 361–365 (cit. on p. 100).

[SDG20] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. "Aspect-controlled neural argument generation." In: *arXiv preprint arXiv:2005.00084* (2020) (cit. on p. 16).

[STV11] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. "On the stratification of multi-label data." In: *Machine Learning and Knowledge Discovery in Databases* (2011), pp. 145–158 (cit. on p. 90).

[Sed12] Philip Sedgwick. "Pearson's correlation coefficient." In: *Bmj* 345 (2012) (cit. on p. 93).

[Sta18] Christian ME Stab. *Argumentative writing support by means of natural language processing*. Gesellschaft für Informatik eV, 2018 (cit. on pp. 18, 21, 35).

[SG14a] 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*. 2014, pp. 1501–1510 (cit. on pp. 10, 22, 38, 39, 61, 66–68, 80, 104, 120).

[SG14b] Christian Stab and Iryna Gurevych. "Identifying argumentative discourse structures in persuasive essays." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 46–56 (cit. on pp. 9, 16, 40, 41, 60).

[SG17a] Christian Stab and Iryna Gurevych. "Parsing argumentation structures in persuasive essays." In: *Computational Linguistics* 43.3 (2017), pp. 619–659 (cit. on pp. 20, 40, 46, 87).

[SG17b] Christian Stab and Iryna Gurevych. “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*. 2017, pp. 980–990 (cit. on pp. 25, 72, 85).

[SGM19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and policy considerations for deep learning in NLP.” In: *arXiv preprint arXiv:1906.02243* (2019) (cit. on p. 110).

[Sun+19] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. “How to fine-tune bert for text classification?” In: *China national conference on Chinese computational linguistics*. Springer. 2019, pp. 194–206 (cit. on p. 88).

[SK17] Piotr Szymański and Tomasz Kajdanowicz. “A Network Perspective on Stratification of Multi-Label Data.” In: *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. Ed. by Luís Torgo, Bartosz Krawczyk, Paula Branco, and Nuno Moniz. Vol. 74. Proceedings of Machine Learning Research. ECML-PKDD, Skopje, Macedonia: PMLR, 2017, pp. 22–35 (cit. on p. 90).

[Tan+16] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. “Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions.” In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 613–624 (cit. on p. 72).

[TP10] Yla R Tausczik and James W Pennebaker. “The psychological meaning of words: LIWC and computerized text analysis methods.” In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54 (cit. on p. 73).

[Tho+22] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. “Lamda: Language models for dialog applications.” In: *arXiv preprint arXiv:2201.08239* (2022) (cit. on p. 110).

[Tol+19] Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. “Automatic Argument Quality Assessment–New Datasets and Methods.” In: *arXiv preprint arXiv:1909.01007* (2019) (cit. on p. 86).

[Tou58] Stephen E Toulmin. *The uses of argument*. Cambridge university press, 1958 (cit. on p. 20).

[Tou03] Stephen E Toulmin. *The uses of argument*. Cambridge university press, Updated Edition, 2003 (cit. on pp. 21, 22).

[UzZ+13] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. "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)*. 2013, pp. 1–9 (cit. on p. 76).

[VE+14] FH Van Eemeren, B Garssen, ECW Krabbe, B Verheij, and JHM Wagemans. *Handbook of argumentation theory. A comprehensive overview of the state of the art*. 2014 (cit. on p. 20).

[VEGSH96] Frans H Van Eemeren, Rob Grootendorst, and Francisca Snoeck Henkemans. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge, Taylor Francis Group, 1996 (cit. on p. 21).

[VEGE04] Frans Van Eemeren, Rob Grootendorst, and Frans H van Eemeren. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004 (cit. on p. 73).

[VSD12] Maria Paz Garcia Villalba and Patrick Saint-Dizier. "Some Facets of Argument Mining for Opinion Analysis." In: *COMMA* 245 (2012), pp. 23–34 (cit. on pp. 78, 82).

[Wac+17a] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. "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: Association for Computational Linguistics, July 2017, pp. 250–255. DOI: 10.18653/v1/P17-2039. URL: https://aclanthology.org/P17-2039 (cit. on pp. 25, 73, 80, 84).

[Wac+17b] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. "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*. 2017, pp. 176–187 (cit. on pp. 25, 26, 81).

[WW20] Henning Wachsmuth and Till Werner. "Intrinsic quality assessment of arguments." In: *arXiv preprint arXiv:2010.12473* (2020) (cit. on pp. 9, 72, 74).

[Wal96] Douglas N Walton. *Argument structure: A pragmatic theory*. University of Toronto Press Toronto, 1996 (cit. on pp. 66, 73).

[Wal97] Douglas Walton. “Appeal to expert opinion. University Park.” In: *PA: Pennsylvania State University Press. Weinstein, EA, Deutschberger, P.(1963). Some dimensions of altercasting. Sociometry* 26 (1997) (cit. on p. 18).

[WR02] Douglas Walton and Chris Reed. “Argumentation schemes and defeasible inferences.” In: *Workshop on computational models of natural argument, 15th european conference on artificial intelligence*. 2002, pp. 11–20 (cit. on pp. 17, 19).

[WR03] Douglas Walton and Chris Reed. “Diagramming, argumentation schemes and critical questions.” In: *Anyone Who Has a View*. Springer, 2003, pp. 195–211 (cit. on pp. 9, 16, 19, 80).

[WRM08] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008 (cit. on pp. 8, 17, 18, 79).

[WMS20] 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 (cit. on pp. 37, 38, 43–45).

[Wan+22] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. “Generalizing to unseen domains: A survey on domain generalization.” In: *IEEE Transactions on Knowledge and Data Engineering* (2022) (cit. on pp. 9, 10, 37, 53, 57).

[Web10] Geoffrey I. Webb. *"Naïve Bayes"*. Ed. by Claude "Sammut and Geoffrey I." Webb. "Boston, MA": "Springer US", "2010", "713–714". ISBN: "978-0-387-30164-8". DOI: "10.1007/978-0-387-30164-8_576". URL: "https://doi.org/10.1007/978-0-387-30164-8_576" (cit. on p. 39).

[WW19] John Woods and Douglas Walton. *Fallacies: selected papers 1972–1982*. Vol. 9. Walter de Gruyter GmbH & Co KG, 2019 (cit. on p. 24).

[WPS21] Zhaofeng Wu, Hao Peng, and Noah A Smith. “Infusing finetuning with semantic dependencies.” In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 226–242 (cit. on p. 109).

[Xu+19] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. “BERT post-training for review reading comprehension and aspect-based sentiment analysis.” In: *arXiv preprint arXiv:1904.02232* (2019) (cit. on p. 88).

[XM12] Huan Xu and Shie Mannor. “Robustness and generalization.” In: *Machine learning* 86.3 (2012), pp. 391–423 (cit. on p. 39).

[XC18] Yumo Xu and Shay B Cohen. "Stock movement prediction from tweets and historical prices." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 1970–1979 (cit. on p. 33).

[Yan+19a] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. "End-to-end open-domain question answering with bertserini." In: *arXiv preprint arXiv:1902.01718* (2019) (cit. on p. 88).

[Yan+19b] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "Xlnet: Generalized autoregressive pretraining for language understanding." In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 86).

[YQX20] Zhen Ye, Yu Qin, and Wei Xu. "Financial Risk Prediction with Multi-Round Q&A Attention Network." In: *IJCAI*. 2020, pp. 4576–4582 (cit. on p. 8).

[Zha+20a] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. "Adversarial attacks on deep-learning models in natural language processing: A survey." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.3 (2020), pp. 1–41 (cit. on pp. 10, 38).

[Zha+20b] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. "Semantics-Aware BERT for Language Understanding." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (2020), pp. 9628–9635. DOI: 10.1609/aaai.v34i05.6510. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6510 (cit. on pp. 56, 109).

[Zha+20c] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. "SG-Net: Syntax-guided machine reading comprehension." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9636–9643 (cit. on p. 109).

[Zhe+19] Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. "Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction." In: *arXiv preprint arXiv:1904.07535* (2019) (cit. on p. 27).

[ZRH20] Shi Zong, Alan Ritter, and Eduard Hovy. "Measuring forecasting skill from text." In: *arXiv preprint arXiv:2006.07425* (2020) (cit. on pp. 31, 73, 74, 76).

[ma+20] Zhiqiang ma, Grace Bang, Chong Wang, and Xiaomo Liu. "Towards Earnings Call and Stock Price Movement." In: *arXiv preprint arXiv:2009.01317* (2020) (cit. on pp. 7, 28, 31, 33, 59).