Exploring Large Language Models in Financial Argument Relation Identification

Yasser Otiefy, Alaa Alhamzeh
University Of Passau
yasser.otiefy@uni-passau.de, alaa.alhamzeh@uni-passau.de

Abstract

In the dynamic landscape of financial analytics, the argumentation within Earnings Conference Calls (ECCs) provides valuable insights for investors and market participants. This paper delves into the automatic relation identification between argument components in this type of data, a poorly studied task in the literature. To tackle this challenge, we empirically examined and analysed a wide range of open-source models, as well as the Generative Pre-trained Transformer GPT-4. On the one hand, our experiments in open-source models spanned general-purpose models, debate-fine-tuned models, and financial-fine-tuned models. On the other hand, we assessed the performance of GPT-4 zero-shot learning on a financial argumentation dataset (FinArg). Our findings show that a smaller open-source model, fine-tuned on relevant data, can perform as a much larger general-purpose one, showing the value of enriching the local embeddings with the semantic context of data. However, GPT-4 demonstrated superior performance with F1-score of 0.81, even with no given samples or shots. In this paper, we detail our data, models and experimental setup. We also provide further performance analysis from different aspects.

Keywords: natural language processing (NLP), argument mining, large language models (LLMs), zero-shot learning, GPT-4, financial domain

1. Introduction

Argumentation plays an indispensable role for financial professionals and market participants. Many investors wait for the quarterly announcements of publicly traded companies to make their investment decisions. The company presents its reports about the last quarter, and makes expectations for the next one, then has to answer professional analysts' questions during a public event of an Earnings Conference Calls (ECCs) (Price et al., 2012). Alhamzeh et al. studied intensively how to mine the arguments of company executives stated during those calls (Alhamzeh et al., 2022b). They revisited the topic and studied how to rank the quality of those arguments in (Alhamzeh, 2023a). They proposed five quality metrics and various types of premises and claims based on interdisciplinary literature. Their study found a considerable link between the argument quality and the relation type (support or attack) between the used premise and the final claim. In other words, an argument that consists of many supporting premises is more likely to be convincing than an argument with many attacking premises. Although discussing the opponent's view is valuable in some situations, the speaker has to state many supporting premises to win.

While this sounds just logical and straightforward, the argument relation detection or classification did not get fair exploration, in the literature, in comparison to other argumentation tasks (e.g., (Reimers et al., 2019; Wachsmuth et al., 2017)). This could be due to its complexity as a Natural Language Inference (NLI) task. However, as we have mentioned, we believe that the potential of solving this task with high accuracy would empower different directions. To clear any possible confusion, on the one hand, the argument relation identification task considers the detection of the relation between given two sentences, so classify them as "related" or "unrelated". In other words, detection if a relation exists between a given premise and claim (the main argument components). While, on the other hand, the argument relation classification task, considers the classification of related premises and claims into a support or attack relation. In our work, we tackle the first identification task, as it is the core stone to structure the argument in the first place.

Furthermore, we focus on the financial use case of argumentation. (Chen et al., 2021) demonstrated, in their book, the urgent need for the automatic mining of arguments in financial narratives and reports. Argument mining considers, mainly, the automatic detection of argument components (premise/claim), argument relations (support/attack), and argument quality assessment.

However, given the fact that financial language has its jargon and particular terms, the language model performance can vary a lot from other domains, even for a simple task like sentiment analysis (Chen et al., 2020). Therefore, the Financial NLP (FinNLP) domain has emerged as an interdisciplinary field, which thus fostered different shared tasks and workshops (e.g., (El-Haj et al., 2018; Shah et al., 2023; Chen et al., 2023a)).

Hence, we have to consider the financial language peculiarities, but also the argumentation discourse nature. Argumentation is proven to be domain-dependent. The structure of arguments can vary a lot between scientific argumentation (Ac-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 119–129
20 May, 2024.

cuosto and Saggion, 2020), legal argumentation (Urchs et al., 2020), or simply web argumentation (Habernal and Gurevych, 2017).

Therefore, with the recent advances in NLP, the need to examine their performance in financial argumentation becomes more urgent. For example, (Al Zubaer et al., 2023) found that a model like Roberta, fine-tuned on the task data, outperform the Generative Pre-trained Transformer (GPT) both versions GPT-3.5[1] and GPT-4 (Achiam et al., 2023) in the legal argument mining area. This raises a critical consideration for each domain. In this paper, we want to assess the performance of large language models in the financial argumentation domain.

In particular, we compare the zero-shot performance of GPT-4, with a wide range of open source Large Language Models (LLMs). We cluster the latter in three categories: general-purpose models (e.g., BERT (Devlin et al., 2019), Vicuna (Zheng et al., 2023)), debate-fine-tuned models (e.g., ArgumentMining-EN-ARI-Debate[2]), financial-fine-tuned models (e.g., FinBert).

The debate-fine-tuned models are fine-tuned on argumentation debate data, while the financial-fine-tuned models are fine-tuned on financial data. Thus, and as our task considers financial argumentation, we aim to inspect the impact of this background data in enriching the model's local embedding.

All in all, the literature lacks a fair exploration of the financial argument relation identification task[3]. We aim, in this study, to bridge this gap. In particular, the contributions of this paper are:

- Empirical study of zero-shot learning and a wide range of outstanding LLMs on Financial Argumentation dataset (FinArg).
- Comparison between the performance of general-purpose, debated-fine-tuned, and financial-fine-tuned LLMs given the nature of this interdisciplinary task.
- To the best of our knowledge, this is the first intensive study to examine recent LLMs on the argument relation identification task.

In Section 2, we navigate the state-of-the-art dedicated to LLMs in argument mining tasks. We overview our data, and methodology in Section 3. Afterward, we exhibit the evaluation results in Section 4. We further discuss and analyze our findings in Section 5. Finally, we conclude our work and open future perspectives in Section 6.

2. Related Work

The exploration of argument mining and text classification has burgeoned with the advent of LLMs. Those models are heavily trained on massive data to learn general language representations. This learned knowledge can be then transformed to downstream domains (or tasks) through the procedure of fine-tuning. This concept made a remarkable revolution in Natural Language Processing (NLP) and helped to solve many challenges, like the need for huge training datasets. However, the behaviour of fine-tuned models on out-of-domain data cannot be completely expected. For example, (McCoy et al., 2019) found that 100 instances of Bert reported performance inconsistency for out-of-domain tests. Similarly, Bert-like models report performance drop in out-of-domain experiments in (Yogatama et al., 2019).

(Ruiz-Dolz et al., 2021) explored BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019a), DistilBERT (Sanh et al., 2019a), and ALBERT (Lan et al., 2019) in identifying argument relations, across various domains. They emphasized the challenge of argument mining due to data scarcity and introduced a comprehensive analysis using the US2016 debate corpus[4] and the Moral Maze corpus[5] and others. The study revealed that different models, especially RoBERTa variants, excel in predicting argument relation on all tested datasets ranging from 0.51 to 0.70 of F1-score, the variation depends on the dataset these models fine-tuned on. This work also emphasizes the potential of other transformer architectures in processing complex argumentative structures.

Since the announcement of GPT-3 in 2020, many studies demonstrated its capability to reach state-of-the-art performance on different NLP tasks without extensive training or fine-tuning. For instance, (Brown et al., 2020) presented a detailed exploration of GPT-3 few-shot learning to generate human-like text, answer questions, translate languages, and other tasks.

The prompt is the main hyperparameter to handle in this scenario. (Liu et al., 2021) provided an exhaustive review of prompt-based learning techniques within NLP. They systematically categorized and evaluated various prompting strategies that leverage the capabilities of pre-trained language models.

In terms of argument mining via LLMs, there have been a couple of research papers that study the

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://huggingface.co/raruidol/ArgumentMining-EN-ARI-Debate

[3] We found only a sub-task in FinArg -1 considering the argument relation classification, which we will address in Section 2

[4] https://corpora.aifdb.org/US2016

[5] https://corpora.aifdb.org/mm2012

power of open-source models fined-tuned to generate semantically rich local embeddings, in comparison to the general OpenAI embeddings. For example, in the legal domain, (Al Zubaer et al., 2023) analyzed the performance of GPT-3.5 and GPT-4 models in classifying argument components (premise/claim) within the European Court of Human Rights dataset. The study found that baseline models (like Large BERT and Roberta) outperform GPT-3.5 and GPT-4, with no significant improvement of GPT-4, over GPT-3.5. Similarly, (Chen et al., 2023b) explored multiple computational argumentation tasks (e.g., claim detection, stance detection) using LLMs in zero-shot and few-shot settings, without any fine-tuning. They found that introducing more samples (longer context) could result in unnecessary information that might negatively affect the performance of smaller models.

From another perspective, (Hinton and Wagemans, 2023) studied how persuasive is AI-generated argumentation. By analyzing the quality of the GPT-3 generator, they concluded that it generated a variety of argument types, but can include fallacies, lacking a real sense of human realization and a cogent argument structure. This raises considerations about the comprehending and reasoning these models can do in argumentation discourses.

In the frame of FinArg-1 shared task (Chen et al., 2023a), argument relation identification task was proposed on a similar dataset derived from (Alhamzeh et al., 2022a), the best team scored 61.50% and 84.86% of macro and weighted F1-score, respectively. Their approach was based on the T5 model (Raffel et al., 2020), fine-tuned using the financial Phrasebank dataset (Malo et al., 2014).

In addition, (Loukas et al., 2023) investigated the use of GPT-3.5 and GPT-4 for few-shot text classification in finance using the Banking77 dataset (Casanueva et al., 2020), demonstrating that conversational LLMs can quickly deliver accurate results and, in some cases, outperform fine-tuned masked language models with fewer examples. However, the cost of subscription-based LLMs may be prohibitive for individuals or smaller organizations. (Li et al., 2023) investigates the efficacy of generically trained LLMs, including ChatGPT and GPT-4, across various financial text analytics tasks, demonstrating their superiority over domain-specific models in many cases but also noting limitations, particularly in tasks requiring deep semantic and structural analysis, this work provides a comprehensive evaluation across eight datasets from five categories of tasks, marking an initial exploration into the capabilities and limitations of LLMs in financial applications.

Hence, and as no consistent superior performance was demonstrated in the recent works on different domains and tasks, we explore in this paper a wide range of LLMs, inspecting their performance on the financial argumentation dataset. Our study is among the first ones to explore the argument relation detection task in a financial narrative.

3. Method

We provide in this section a detailed overview of the data, models, and our experimental setup.

3.1. Data

We conducted our experiments on the Financial Argumentation dataset *FinArg*, which was collected and annotated by (Alhamzeh et al., 2022b; Alhamzeh, 2023b). This data is publicly available[6], and covers the quarterly earnings conference calls of major corporations (Amazon, Apple, Microsoft, and Facebook[7]) spanning from 2015 to 2019.

The annotation of this data encompasses the following labels: *premise*, *claim*, *non-arg* on the sentence level, as well as *support/attack* label on the relation between related premises and claims. Therefore, and to be able to solve the relation identification problem, we had to deduce the unrelated relation examples from the data. Subsequently, we construct our data as follows:

- **Positive Sampling:** We concatenate each claim with every single corresponding premise using [SEP] token (i.e., claim [SEP] premise), and we label it with class ‘1’, signifying a related pair. This outcome in about 5K samples generated from 2200 arguments.
- **Negative Sampling:** We pair the unrelated claim-premise pairs and label each with class ‘0’. By this, we got about 1M possible pairs.
- **Data Balancing:** To keep the data balanced, we randomly selected 5K negative samples.

Hence, our problem is a binary classification task, on a balanced dataset. We have approx. 10K data samples formatted as the following:

- **Input –> {Claim} [SEP] {Premise}**
- **Output –> "1" or "0"**

3.2. Models

In this section, we elaborate on our models and experimental setup. We have examined two families of state-of-the-art large language models. On the first hand, fine-tuned models from Huggingface[8],

[6] FinArg Dataset
[7] Currently Known as Meta
[8] https://huggingface.co

and on the other hand, GPT language model from OpenAI[9]. This setting allows us to inspect the impact of the fine-tuning phase on the output in comparison to generative models where the prompt plays a considerable role.

3.2.1. Fine-tuned Large Language Models

To investigate the potential of open-source LLMs in argument relation identification, we examine in our study three categories of models, based on their training data, and intended application. This classification enables a focused analysis of each model's performance, especially in tasks that align with their customized training. We provide in the following an overview of those categories, and the examined models corresponding to each.

1. **General-purpose models:** This category encompasses original models that have been trained on general domain-agnostic data. These models are designed to perform a variety of natural language understanding tasks across different domains due to their diverse training backgrounds. Our used models from this category include:
   - *Bert-base-uncased* (Devlin et al., 2019)
   - *Roberta-base* (Liu et al., 2019b)
   - *Distilbert-base-uncased* (Sanh et al., 2019b)
   - *Bloom (560m,1b,7b)* (Workshop et al., 2022)
   - *BloomZ* (Muennighoff et al., 2022)
   - *LLaMa-2-7B-Guanaco-QLoRA-GPTQ*[10] a fine-tuned version of Llama 2 (Touvron et al., 2023)
   - *Vicuna*: is a chat assistant trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. We test two versions *(Vicuna13bv1.5 and Vicuna-13b_rm_oasst_hh*[11]) (Zheng et al., 2023)
   - *GPT4-X-Alpaca*[12] a finetuned on GPT4's responses, for 3 epochs of a base model Alpaca (Taori et al., 2023)
2. **Debate-fine-tuned models:** Models in this category have been specifically fine-tuned on datasets featuring argumentative structures derived from debate content, which can be related to finance. They are optimized to discern and process argumentative nuances, making them well-suited for applications of argument mining. We include in this category:
   - *ArgumentMining-EN-ARI-Debate*, *ArgumentMining-EN-AC-Essay-Fin*, *ArgumentMining-EN-AC-Financial*, *ArgumentMining-EN-CN-ARI-Essay-Fin*[13]: All adopted from (Ruiz-Dolz et al., 2021), as fine-tuned versions of (Conneau et al., 2019) on different datasets such as US2016-test, MM2012, Bank, Money and others. For more details about those models, please refer to (Ruiz-Dolz et al., 2021).
   - *Roberta-argument*[14] trained on 25k heterogeneous manually annotated sentences by (Stab et al., 2018) and *Roberta-base-150T-argumentative-sentence-detector*[15]: A fine-tuned version of RoBerta (Liu et al., 2019b) using FS150T-Corpus dataset by (Schiller et al., 2022).
3. **Financial-fine-tuned models:** Our third category consists of models that have been fine-tuned with financial datasets, aiming to address classification challenges pertinent to the financial sector. These models leverage financial discourse and numeric data to provide insights specific to financial contexts. Namely:
   - *Finbert* (Araci, 2019) involves enhancing the BERT language model specifically for the finance sector. This is achieved by training it on a substantial corpus of financial documents, subsequently refining its capabilities for classifying financial sentiment. For this fine-tuning process, the Financial PhraseBank, created by (Malo et al., 2014), is employed.
   - *Finbert-tone-finetuned-finance-topic-classification* (Hazourli, 2022): Fine-tuned version on sentiment analysis task on Financial PhraseBank by (Malo et al., 2014).
   - *Deberta-v3-base-finetuned-finance-text-classification*[16]: Fine-tuned version of Deberta (He et al., 2021) tuned on financial-classification dataset[17].

[9] https://openai.com

[10] https://huggingface.co/TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ

[11] https://huggingface.co/reciprocate/vicuna-13b_rm_oasst-hh

[12] https://huggingface.co/chavinlo/gpt4-x-alpaca

[13] https://huggingface.co/raruidol

[14] https://huggingface.co/chkla/roberta-argument

[15] https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector

[16] https://huggingface.co/nickmuchi/deberta-v3-base-finetuned-finance-text-classification

[17] https://huggingface.co/datasets/nickmuchi/financial-classification

- *Roberta-Earning-Call-Transcript-Classification*[18]: Fine-tuned model from the base model RoBerta (Liu et al., 2019b) tuned on extracted a decade's worth of earnings call transcripts for 10 corporations, including Apple, Google, Microsoft, Nvidia, Amazon, Intel, Cisco, and others.

In all these categories, we conduct 5-fold cross-validation, with hyperparameter optimization as follows:

- Learning rate (2e^{-5}, 3e^{-5}, 5e^{-5})
- Maximum length of the tokenizer (64, 128, 256)
- Number of epochs (ranging from 2 to 5)

Please note that all fine-tuned models are trained on 2 x NVIDIA A100 80GB GPUs using Pytorch Lightening and HuggingFace frameworks with global seed 42.

3.2.2. GPT-4 Zero-Shot Learning

In our experiments, we explore the capability of the *GPT-4* model (Achiam et al., 2023) to detect the relation between a given claim and premise, using zero-shot learning (Xian et al., 2018).

Zero-shot learning refers to the model's ability to understand and perform tasks without the need for a specific training dataset tailored to that task. Recently, it has shown a very competent performance in various NLP tasks (Wei et al., 2021; Brown et al., 2020).

Prompt Design As prompting has not been yet explored in the task of financial argument relation detection, and due to budget constraints, we chose to follow a basic hand-crafted prompt. This is also justified by the fact that the prompt has a significant impact in few-shot learning where choosing the number of shots, and choosing the example(s) play a crucial role, also this is impacted by budget constraints whereas we apply a zero-shot experiment.

Therefore, we decided to follow a straightforward approach that gathers the context and the instruction to the model (Brown et al., 2020). Obviously, we consider carefully OpenAI recommendations and prompt guide[19] as well as the prompt engineering guide[20].

Since we aim to classify the relation between a given claim and premise as either *Related* or *Unrelated*, we formulate our prompt to clarify those two explicitly and then ask for the output class, as shown in the function `generate_messages` in the following:

```
def generate_messages(claim, premise):
    messages = [
        {"role": "system", "content":
            "You are a helpful
            assistant. Given the
            following claim and
            premise, please classify
            the relation between them
            as either Related or
            Unrelated. Please only
            generate one of the two
            labels."},
        {"role": "user", "content":
            f"Claim: {claim}"},
        {"role": "user", "content":
            f"Premise: {premise}"},
    ]
    return messages
```

This function encapsulates the interaction pattern with the model, where the model is first instructed about its role and the task's objective. Following this, the claim and premise are presented for classification.

Post-Processing of GPT-4 Output Following the interaction with the *GPT-4* model (Achiam et al., 2023), a crucial step is required to accurately extract the classification labels. The model responses are encapsulated within structured formats either as content within the interaction messages or through explicit function call objects which require systematic extraction processes to discern the relation classification between claims and premises. In other words, we had to check the extracted class label, to ensure it aligns with the expected output format and classification options ('Related' or 'Unrelated'). In some cases, the model responds by undefined class, then we have to extract it from the function call[21] output, if it does not exist in both response and function call response we label the sentence with "Unrelated" since this is the safe solution.

4. Results

In our comprehensive evaluation of argument relation identification, we explored a wide spectrum of fine-tuned Large Language Models (LLMs) alongside the innovative zero-shot learning capabilities of *GPT-4*, unveiling a fascinating landscape of performance across models tailored for General-purpose, Debate-fine-tuned, and Financial-fine-tuned tasks.

[18] https://huggingface.co/NLPScholars/Roberta-Earning-Call-Transcript-Classification

[19] https://platform.openai.com/docs/guides

[20] https://www.promptingguide.ai/techniques/zeroshot

[21] https://platform.openai.com/docs/guides/function-calling

To have comparable results, we train the fine-tuned models in a cross-validation approach, where each part of the data is a test set at some fold. We then consider all data (all possible test sets) as the test data for *GPT-4*. Therefore, we report in Table 1 the average performance of the fine-tuned models along with the standard deviation, while we report in Table 2 the outcomes of *GPT-4* considering all the data.

Our results show that *GPT-4* was the most efficient performer by a macro F1-score of 0.81, confirming its ability to grasp the nuances of argumentative relations without explicit task-specific training.

However, among the fine-tuned models, *Vicuna-13b_rm_oasst_hh*, and *ArgumentMining-EN-ARI-Debate* showed a good performance with a mean macro F1 Score of 0.751. Despite the huge difference in the number of parameters between those two models, the latter behaved closely to Vicuna, only by having it already fine-tuned on debate data. This reflects the custom data impact on handling domain-specific argumentation. Yet, both models of *ArgumentMining-EN-CN-ARI-Essay-Fin* and *ArgumentMining-EN-AC-Financial* exhibited poor recognition of the argument relation.

In the series of Bloom models, the version of *Bloom 7b* parameters achieved a mean F1-score of 0.65, whereas a random guess behaviour was observed by *Bloom 560 m*, *Bloom 1b*, and *Bloomz 7b*. Similarly, FinBert, llama-2, Bert, and Alpaca showed weak efficiency. At the bottom of the list, lags *Roberta-Earning-Call-Transcript-Classification*, with an F1-score of 0.371, indicating a potential misalignment with the dataset's characteristics or the need for further tuning.

Our zero-shot learning experiment, which was conducted with *GPT-4*, is detailed in Table 2. It reveals *GPT-4* robust classification ability, with a precision of 0.85 for "Related", and 0.77 for "Unrelated" classes, reflecting a balanced understanding of both relationship types. This performance is further encapsulated in the precision-recall balance, with *GPT-4* favouring recall for "Unrelated" (0.87) over "Related" (0.75), suggesting a slight inclination towards conservatively identifying unrelated pairs to mitigate the risk of false positives in argumentative contexts.

The aggregate analysis does not only highlight the superior adaptability and understanding of *GPT-4* in zero-shot learning scenarios but also points to significant variations in the effectiveness of fine-tuned models across different categories. These distinctions underline the importance of model selection tailored to the specific characteristics of the task at hand, where the data domain and the classification task's nature critically influence model performance. The breadth of models evaluated demonstrates a spectrum of capabilities, from the comprehension exhibited by *GPT-4* to the more domain-specific insights offered by models like *Vicuna 13b*, and *ArgumentMining-EN-ARI-Debate*.

5. Discussions

In this section, we will discuss the analysis of hyperparameters, also we will spotlight the models that significantly outperformed the other models and attempt to justify these gaps. Since our data is balanced, we will focus on discussing the mean macro F1-score as it captures the harmonic mean of precision and recall.

The variability in performance as indicated by the standard deviation from the 5-fold cross-validation process as shown in Table 1 reveals insights into model stability. In general, models showed low standard deviations, suggesting consistent performance across different data folds and thus, greater reliability in practical applications.

The impact of model size on the F1-score in Figure 1 was evident from the visual data. While larger models generally achieved higher F1-score, indicating better generalization, the increase of model size did not always correlate with proportional improvements of results. This suggests a point of diminishing returns, where additional model complexity yields minor improvements at a significant computational cost. However, some models with a small number of parameters achieved relatively good performance. Potential reasons are the domain of the data those models used for tuning and also the task that those models tuned on, when possibly similar to our task, argument relation identification.

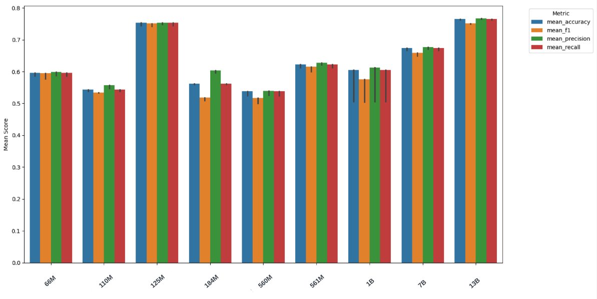

Figure 1: A grouped bar chart displaying the comparison of four metrics mean (accuracy, F1 score, precision, and recall) across models of various sizes.

Figure 2 indicates the performance of the three categories of open-source models we have experimented with. It reflects that Debate-fine-tuned and General-purpose models have a comparable mean macro F1-score, outperforming the Financial-fine-tuned models. This may suggest that general reasoning knowledge learned in debate-fine-

Model	Accuracy	F1-score	Precision	Recall	Model Type
Vicuna-13b_rm_oasst-hh	0.764 ± 0.05	**0.751 ± 0.05**	0.767 ± 0.05	0.764 ± 0.05	General-Purpose Models
Vicuna-13b-v1.5	0.762 ± 0.05	0.750 ± 0.05	0.762 ± 0.05	0.762 ± 0.05	
Bloom-7b1	0.675 ± 0.04	0.659 ± 0.06	0.677 ± 0.04	0.674 ± 0.04	
Bloom-1b1	0.567 ± 0.04	0.549 ± 0.05	0.572 ± 0.04	0.567 ± 0.04	
Bloomz-7b1	0.567 ± 0.02	0.534 ± 0.03	0.573 ± 0.02	0.567 ± 0.02	
Bloom-560m	0.531 ± 0.02	0.507 ± 0.03	0.530 ± 0.02	0.531 ± 0.02	
Bert-base-uncased	0.532 ± 0.01	0.503 ± 0.03	0.541 ± 0.02	0.532 ± 0.01	
GPT4-x-Alpaca	0.558 ± 0.04	0.536 ± 0.04	0.561 ± 0.04	0.558 ± 0.04	
LLaMa-2-7B-Guanaco-QLoRA-GPTQ	0.517 ± 0.01	0.468 ± 0.06	0.504 ± 0.09	0.517 ± 0.01	
Roberta-base	0.547 ± 0.03	0.479 ± 0.09	0.563 ± 0.13	0.547 ± 0.03	
ArgumentMining-EN-ARI-Debate	0.753 ± 0.02	**0.751 ± 0.02**	0.753 ± 0.01	0.753 ± 0.02	Debate-fine-tuned Models
ArgumentMining-EN-AC-Essay-Fin	0.622 ± 0.04	0.615 ± 0.04	0.627 ± 0.02	0.622 ± 0.02	
Roberta-base-150T-argumentative-sentence-detector	0.578 ± 0.01	0.569 ± 0.01	0.584 ± 0.02	0.578 ± 0.02	
ArgumentMining-EN-CN-ARI-Essay-Fin	0.532 ± 0.01	0.492 ± 0.07	0.540 ± 0.06	0.532 ± 0.01	
ArgumentMining-EN-AC-Financial	0.530 ± 0.02	0.480 ± 0.08	0.536 ± 0.09	0.530 ± 0.02	
FinancialBERT-Sentiment-Analysis	0.518 ± 0.02	**0.514 ± 0.02**	0.518 ± 0.02	0.518 ± 0.02	Financial-fine-tuned Models
Roberta-Earning-Call-Transcript-Classification	0.503 ± 0.01	0.371 ± 0.07	0.359 ± 0.14	0.503 ± 0.01	
Finbert	0.516 ± 0.02	0.507 ± 0.03	0.517 ± 0.02	0.516 ± 0.02	
Deberta-v3-base-finetuned-finance-text-classification	0.554 ± 0.01	0.505 ± 0.03	0.589 ± 0.02	0.554 ± 0.01	

Table 1: Classification performance metrics of LLMs on argument relation identification using 5-fold cross-validation. All models reported here are fine-tuned for 5 epochs, except Bloomz-7b1, for 2 epochs. The learning rate for all models is 5e^{-5}

Class	Precision	Recall	F1-score	Support
Related	0.85	0.75	0.79	4899
Unrelated	0.77	0.87	0.82	4899
Accuracy			0.81	9798
Macro Avg	0.81	0.81	0.81	9798
Weighted Avg	0.81	0.81	0.81	9798

Table 2: Classification performance metrics of *GPT-4* zero-shot learning

tuned models is more valuable than the financial background knowledge learned in the Financial-fine-tuned models. Yet, the performance between Debate-fine-tuned models and General-Purpose Models is comparable, which could rely on the size of the latter. Therefore, we suggest examining smaller LLMs for a low tuning cost before looking for huger models, especially in a small dataset setting.

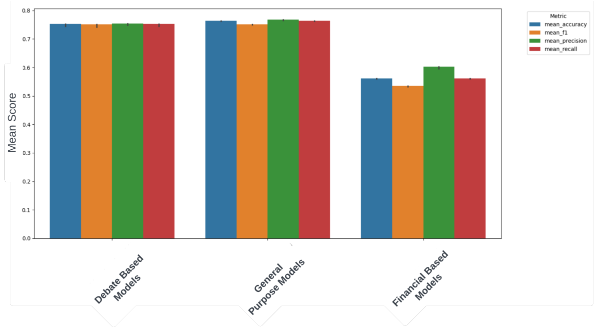

Figure 2: Performance among the three categories of fine-tuned models (Debate-fine-tuned, General-purpose, Financial-fine-tuned)

Figure 3, and the Pearson correlation heat map presented in Figure 4 provide an understanding of the relationship between hyperparameters and F1-score. Certain hyperparameters such as epochs and learning rate showed positive correlations with the F1-score. Potentially, since we give the model the chance to distil the pattern of our data, which means the more epochs we give to the model to train, the better the model learns.

Hyperparameters like maximum input length (max length), did not exhibit a very strong relationship with mean F1-score since most of the data points, as shown in Figure 5, are less than the smallest value of the max length hyperparameter ranging from (64 to 256) and the frequency of the examples that has 64 tokens or less is dominant. However, the correlation still exists which means the longer the sentence is fed to the model without truncation, the better performance the model achieves. However, a complex interplay between these hyperparameters requires careful tuning to optimize performance.

We also have noticed that the standard deviation, in general, is small which means the consistent performance of such models with low standard deviation, however, some models have a slightly larger standard deviation such as Roberta-base and *ArgumentMining-EN-AC-Financial*, One of the reasons could be the type of data these models fine-tuned on which made those models overfitted and stuck in a local minimum because of such past fine-tuning.

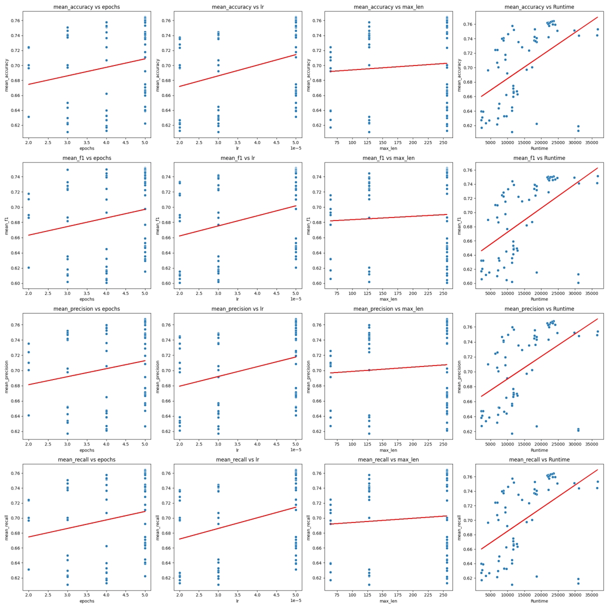

Figure 3: Correlation between hyperparameters (epochs, learning rate, input max length, runtime) and the performance metrics of fine-tuned models (accuracy, F1-score, precision, recall)

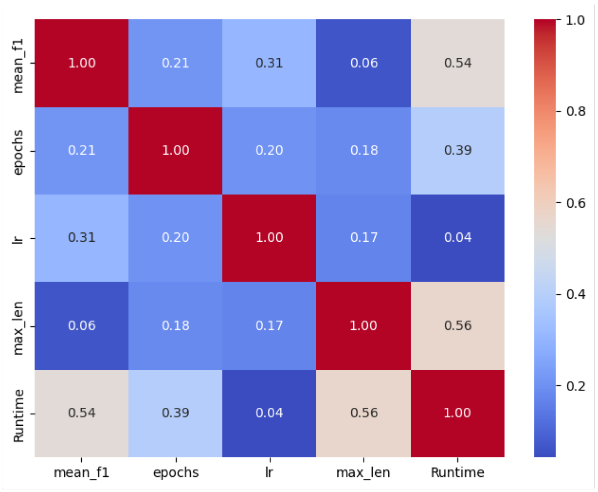

Figure 4: The heat map shows that learning rate and runtime, maximum input length and epochs correlation with mean F1-score.

6. Conclusion

The automatic mining of arguments (components and relations) has become an essential tool for multiple applications like assisted writing, fact-checking, search engines, law, and decision-making aid systems. In this paper, we investigated argument mining in financial texts, In particular, the task of relation detection between given two sentences (potential argument components) within the context of earnings conference calls.

Our experimental study encompasses a wide range of LLMs, including *GPT-4*, debate-fine-tuned models, and financial-fine-tuned models. The performance of open-source models ranged from 0.37 to 0.75 in terms of F1-score, while *GPT-4* zero-shot learning achieved 0.81. This superior performance of *GPT-4* highlights its potential to adapt to complex language understanding tasks, without any further training. Moreover, we believe that this outcome can be significantly improved with few-shot learning, or exploring other prompting techniques in future work.

In closing, our study contributes to the literature of argument mining in the financial domain by providing a comprehensive evaluation of various LLMs and illustrating the potential of zero-shot learning in understanding the nuances of financial discourse.

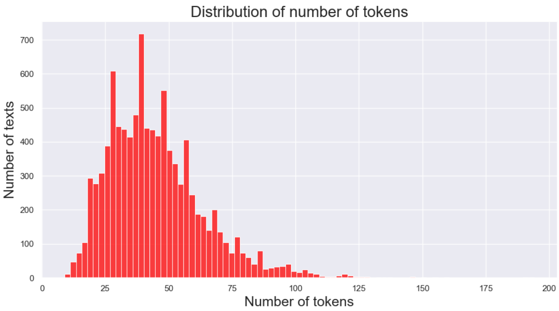

Figure 5: Distribution of sentence length

7. Bibliographical References

Pablo Accuosto and Horacio Saggion. 2020. Mining arguments in scientific abstracts with discourse-level embeddings. *Data Knowledge Engineering*, 129:101840.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6.

Alaa Alhamzeh. 2023a. Financial argument quality assessment in earnings conference calls. In *International Conference on Database and Expert Systems Applications*, pages 65–81. Springer.

Alaa Alhamzeh. 2023b. *Language Reasoning by means of Argument Mining and Argument Quality*. Ph.D. thesis, Universität Passau.

Alaa Alhamzeh, Előd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2022a. *Empirical Study of the Model Generalization for Argument Mining in Cross-Domain and Cross-Topic Settings*, pages 103–126. Springer Berlin Heidelberg, Berlin, Heidelberg.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022b. 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)*, pages 163–169.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

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

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

Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023a. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023b. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj, Paul Rayson, and Andrew Moore. 2018. The first financial narrative processing workshop (fnp 2018). In *Proceedings of the LREC 2018 Workshop*.

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

Ahmed Hazourli. 2022. Financialbert-a pretrained language model for financial text mining. *Research Gate*, 2.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Martin Hinton and Jean HM Wagemans. 2023. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument & Computation*, (Preprint):1–16.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with ChatGPT: Few-shot text classification for finance. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao. -.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

R Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

R. Ruiz-Dolz, J. Alemany, S. Barbera, and A. Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(06):62–70.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. On the effect of sample and topic sizes for argument mining datasets. *arXiv preprint arXiv:2205.11472*.

Sameena Shah, Xiaodan Zhu, Wenhu Chen, Manling Li, Armineh Nourbakhsh, Xiaomo Liu, Zhiqiang Ma, Charese Smiley, Yulong Pei, and Akshat Gupta. 2023. Knowledge discovery from unstructured data in financial services (kdf) workshop. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3464–3467.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2020. Towards classifying parts of german legal writing styles in german legal judgments. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 451–454. IEEE.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. 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*, pages 176–187.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.