NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002044

SCaLAR IT at the NTCIR-18 FinArg-2: Temporal Inference of Financial Arguments

Sai Saketh Nandam[1]
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
nandamsaisaketh.242it021@nitk.edu.in

Charan Srinivas Kumar Reddy Dasari[2]
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
dcharanreddy.242it021@nitk.edu.in

Anand Kumar Madasamy[3]
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
m_anandkumar@nitk.edu.in

Abstract

The SCaLAR IT team participated in the Detection of Argument Temporal References subtask of the NTCIR-18 FinArg-2 Task. This paper presents our approach to solving the classification of financial arguments based on temporal references. We explored multiple architectures combining a BERT-based model with knowledge-based and temporal feature extraction techniques. To improve the performance,integrated BERT with TF-IDF based temporal features were extracted using STANZA and BERT embeddings to enhance temporal reference detection. Our first model BERTForSequenceClassifier achieves the Micro F1 score of 70.24% and Macro F1 score of 67.85% outperforming most approaches of other teams. However incorporating additional temporal features improved the Macro F1 score, indicating better performance across all classes. We analyze the effectiveness of different feature representations in our research.

Keywords

Pre-trained, Argument, BERT, Stanza, Knowledge base

Team Name

SCaLAR IT

Subtasks

Detection of Argument Temporal References (English)

1 Introduction

The Scalar team participated in the Argument Temporal Reference Detection subtask under the FinArg-2 task at NTCIR-18 [6]. This subtask utilizes the Earnings Conference Calls (ECC) dataset, which aims to classify financial arguments based on the presence and type of temporal references, distinguishing between arguments with no time reference, short past references, and long past references. By identifying such temporal cues, this task enhances automated financial information processing and supports more accurate decision-making models.

An earnings call is a quarterly event during which publicly traded companies review their performance from the previous quarter and provide forecasts for the upcoming quarter. During the Q&A sessions of these earnings conference calls, company managers respond to questions posed by financial analysts and other market participants. These discussions play a significant role in shaping perceptions and can have a notable impact on financial markets. Studies show that this part of the call is often the most informative, providing deeper insights into the company's strategies, risks, and opportunities. An argument consists of one claim and one premise supporting it.

Examining the temporal aspects of financial information and determining its period of impact are critical research questions in the stock market. Temporal inference in financial statements is a crucial task in understanding and interpreting the timing of financial events, forecasts, and discussions. Financial texts, such as reports, news articles, and analyst briefs, often contain explicit and implicit references to time, such as specific years, quarters, or temporal phrases (e.g., "next quarter," "last year"). The ability to identify and classify these temporal references plays a significant role in automating the analysis of financial data, making it easier for analysts to track and compare timelines of financial events.

In the broader context, earnings calls have been increasingly recognized as a valuable resource for predictive financial modeling. These calls provide not only structured data in the form of reported financial results but also unstructured textual data, such as management's qualitative commentary. Leveraging Natural Language Processing (NLP) to analyze these texts enables researchers to extract actionable insights and detect patterns that are otherwise hidden.

2 Related Works

Chen et al.(2021)[5] focuses on understanding the rationales provided by amateur investors in financial contexts. This study evaluates their reasoning, which often consists of temporal elements like past performance or projections. The methods and results align with analyzing temporal markers in financial arguments to analyze their impact on investment decisions. Chen et al.(2018)[4] investigates the role of numeral understanding in financial tweets for fine-grained crowd-based forecasting. Temporal references in tweets, such as "next quarter" or "last year," are critical for extracting actionable insights. The study highlights the importance of proper numeral and temporal interpretation within financial data.

UzZaman et al. (2013)[9] developed assesment guidelines for temporal expressions, relations, and events. The TempEval-3 tasks constitute the basis of the retrieval and inference time-sensitive information, creating procedures that are scalable to financial contexts for temporal inference. Palmieri (2023)[1] and related works

Claim:
But at this point in time, we see that Q2 is the toughest compare.
Premise:
If you recall, we were heavily supply constrained throughout the whole of Q1 and so some of that demand moved into Q2. Plus we're in an environment now that is dramatically different from a macroeconomic point of view than last Q2, from a currency point of view, from the level of which we've had to adjust pricing in several of these markets, and sort pf the overall Indiscernible in virtually every country in the world., because of the year ago quarter also had catch up in it from Q1.
Referenced Year: 2016
Referenced Quarter: Q1
Label: 1 (Long Past)

Figure 1: An example from the dataset, showcasing a claim-premise pair along with the referenced year, quarter, and label assignment.

emphasize temporal reasoning within business communication during earnings calls. They highlight how the timing plays a role with information (e.g., past and future-oriented) in the context of driving market sentiment and risk assumptions.

Stanza [8] is a deep learning-based NLP toolkit designed for both syntactic and semantic processing. It is used for sentence syntax understanding as well as in the identification of temporal markers, such as time-dependent verbs and date-related entities.

3 Dataset Overview

The Earnings Conference Calls(ECCs)[2] is typically well-organized, with transcripts containing claims, premises, and time markers. It contains 600 records for training and 150 records for validation. Each record includes argument's claim and premise sourced from the years 2015 to 2019 [3]. Additionally, the dataset contains ground truth labels alongside the year and financial quarter in which each argument was made or claimed. It contains several key components that contribute to the classification process. The claim text serves as the primary argument or statement to be classified and is essential for feature extraction. Figure 1 presents a sample instance from the dataset.

The classification labels for the task are:

- 0: No time reference
- 1: Long past (more than half a year)
- 2: Short past (less than half a year): during this quarter or up to 2 quarters.

The primary assumption is that we concatenated the claim and premise to form a single unified argument, which was used as input for all the models. This approach ensured that the models could analyze the combined contextual and supportive information provided by the premise alongside the claim.

4 Methodology

In this section, we present the proposed methodology for detecting temporal references in financial arguments. We describe the integration of temporal feature with transformer based model to enhance model performance.

4.1 Knowledge based Approach

This approach was developed to classify arguments based on temporal references, especially focusing on the years and quarters mentioned in the text. Temporal references, such as financial quarters (e.g., Q1, Q2, Q3, Q4) and four-digit year references were identified using regular expressions. Each quarter was assigned to its numerical time index Q1: 1, Q2: 2, Q3: 3, and Q4: 4. The temporal distance was computed by measuring the maximum duration between the extracted temporal references and the referenced time stamps of the argument.The classification was based on these calculated differences:

- If no temporal references were extracted from the text, the argument was labeled 0 (no temporal reference) and temporal distance is 0.
- If only quarter differences were present , the argument was labeled 2 if quarter difference is less than or equeal to 2;otherwise, it was labeled 1. The temporal distance is the quarter difference.
- If only year differences were present, arguments were labeled 2 if the year difference was zero;otherwise, it was labelled 1. The temporal distance is computed as the year difference multiplied by 4 quarters.
- If both year and quarter differences were present, temporal distance is computed as (year difference * 4) + (quarter difference). The argument was labeled 2, if the temporal distance is less than or equeal to 2;otherwise, it was labeled 1.

4.2 Transformer based Approach

BERT. In this approach, we used a BERT pre-trained transformer[7] for argument classification. The input text was passed to Bert Tokenizer for data preprocessing, which converts text into token IDs. Tokenizer adds a special token such as [CLS] and [SEP] indicating the start and end of each input sequence. Padding and truncation were applied to adjust all the inputs to a fixed length before passing to the model.

We fine-tuned a BERT-based sequence classification model for our multi-class classification task which predicts the probability of target labels 0, 1, and 2.

BERT + knowledge based. The BERT model captures the deep contextual representation of the argument and the knowledge-based approach explicitly focuses only on the temporal references without considering the context. We propose a hybrid model to leverage the strengths of knowledge-based and deep learning techniques. First, The BERT tokenizer processes the argument and outputs a rich semantic embedding using the [CLS] token. Next, similar to a knowledge-based system we extract the temporal references within the argument and compute two numerical features: a reference presence and temporal distance. Reference presence is a binary number if 1 indicates that temporal reference is present else

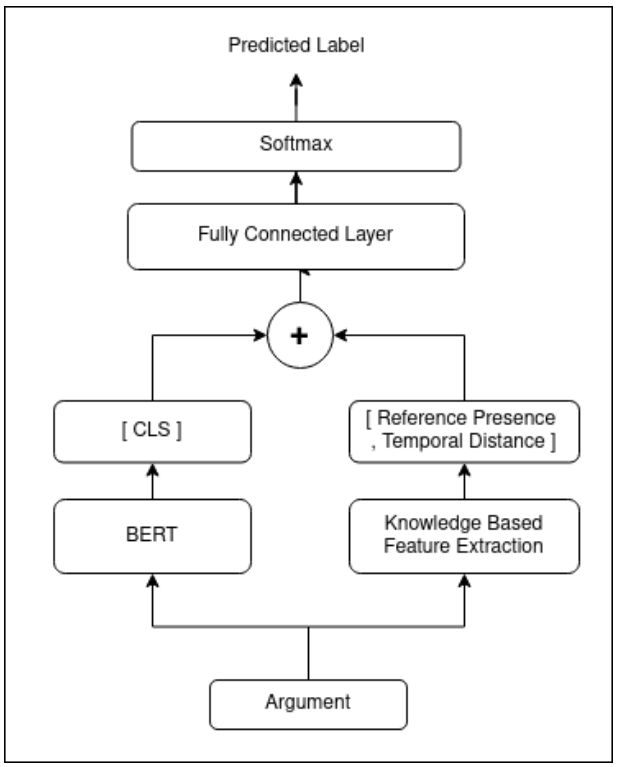

Figure 2: Model architecture for the BERT + knowledge-base.

0 and a temporal distance represents the maximum temporal gap in terms of quarters between the referenced time and the identified temporal mentions in the argument. These features give the temporal information. The extracted features from the knowledge-based approach are concatenated with the BERT [CLS] embedding. This argument representation is then fed into a fully connected classification layer, which predicts the probabilities of the target label. Figure 2 shows the overall architecture of this approach.

BERT + Stanza. In this approach, we integrate temporal information extraction using Stanza to enhance the classification of financial arguments. The first step involves extracting temporal expressions from the argument using Stanza's Named Entity recognition. Apply the the Stanza pipeline to identify and extract entities labeled as 'DATE'. These extracted date entities are concatenated into a single string and considered as temporal text. This ensures that all temporal references in the argument are explicitly captured. In the second step, the temporal text is preprocessed by removing stopwords, lowercasing, and normalizing special characters. Then transform it into a feature vector using the TF-IDF vectorizer, which captures the frequency and importance of temporal terms within the dataset.

These TF-IDF vector serve as an additional input feature, which is later concatenated with BERT embeddings derived from the argument. This combined feature vector is passed through a fully connected classification layer, which outputs the probability distribution over the target classes. By combining deep contextual embeddings from BERT with structured temporal information, the model takes advantage of both semantic meaning and explicit temporal cues for classification. Figure 3 shows the overall architecture of this approach.

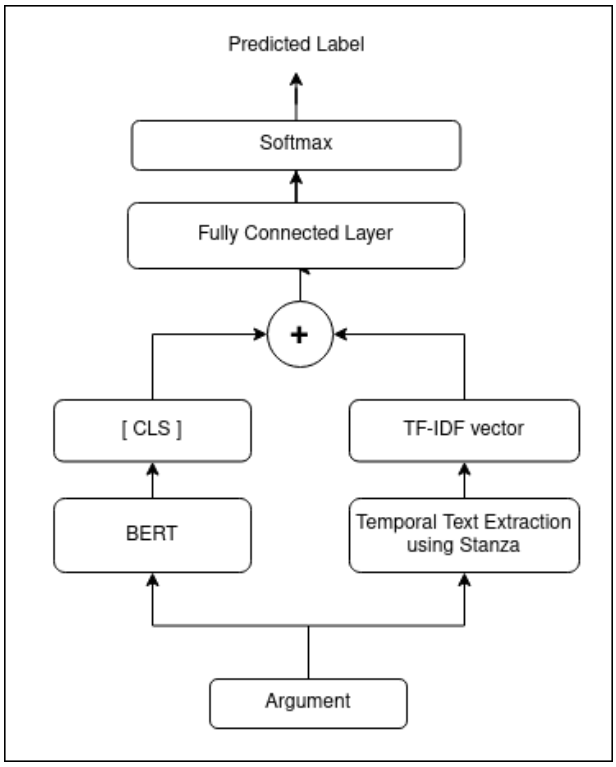

Figure 3: Model architecture for the BERT + Stanza.

In addition to TF-IDF vectorization, we also experimented with using BERT tokenizer to process the extracted temporal text. Instead of generating a sparse TF-IDF vector, the temporal text was tokenized using the BERT subword tokenization and converted to dense embeddings. Then these embeddings were concatenated with the [CLS] embedding of the argument, providing a richer semantic representation of temporal references.

5 Experimental Setup

In this study, we experimented different approaches for financial argument classification. The FinArg-2 provided both training set and validation set. We further divided validation set into two equeal parts: one for validation and other for testing. Dataset Statistics are shown in table 2.

In all our experiments, we used bert-base-uncased version model. For our first approach, we fine-tuned the BERTForSequenceClassification model from Hugging Face's transformers library. In the third approach, The TF-IDF feature size was determined based on the training data, ensuring that the temporal representation was compact and informative. In SCaLAR IT_ECC_3, the model concatenates the 768-dimensional [CLS] embedding from BERT with a TF-IDF feature vector of size 133. In SCaLAR IT_ECC_4, the model concatenates the 768-dimensional [CLS] embedding of argument

Table 1: Model Performance Comparison on our Test Set

Model Name	Architecture	Embedding	Micro F1(%)	Macro F1(%)	Weighted F1
SCaLAR IT_ECC_1	BERTForSequenceClassifier	768	**79.00**	75.00	**79.00**
SCaLAR IT_ECC_2	BERT + knowledge Based	768 + 2	76.00	75.33	77.08
SCaLAR IT_ECC_3	BERT + Stanza (TF-IDF)	768 + 133	77.33	**75.69**	77.69
SCaLAR IT_ECC_4	BERT + Stanza (BERT Tokenizer)	768 + 768	69.33	68.11	70.38

Table 2: Dataset Statistics

Split	Label 0	Label 1	Label 2	Total
Train	301	123	176	600
Validation	36	20	19	75
Test (Ours)	39	15	21	75

with 768-dimensional [CLS] embedding of temporal text, which is extracted using Stanza. The hyperparameter settings are detailed in Table 3.

Table 3: Hyperparameter Settings

Parameter	Value
Learning rate	1e-5, 2e-5
Optimizer	AdamW
Loss function	Cross Entropy
Epochs	20

5.1 Experimental Results and Analysis

The results shown in Table 1 demonstrate the performance comparision of the experimented models in terms of Mircro F1, Macro F1, and Weighted-F1. The BERTForSequenceClassifier(SCaLAR IT_ECC_1) achieved the highest Micro F1 and Weighted F1 scores of 79.00%, making it the most balanced model in terms of overall accuracy across classes. The SCaLAR IT_ECC_2 showed moderate improvement over with 75.33% Micro F1. Among Stanza-based models, SCaLAR IT_ECC_3 achieved the highest Macro F1 score of 75.69%, performed well in class-wise evaluation. However, SCaLAR IT_ECC_4 performed the worst among all models with Micro F1 score of 69.33% and Macro F1 score of 68.11%.

The use of the full 768-dimensional dense BERT embeddings did not lead to performance improvement, suggesting that TF-IDF's sparse representations capture temporal features more effectively than dense embeddings. Overall, the results suggests that while BERT alone remains strong baseline for classification task, integrating TF-IDF based temporal features improves class-wise balance. However, using BERT embeddings for both argument and temporal text doesn't yield better performance.

Upon analyzing the output labels generated by different models, it was observed that the BERT-based model primarily classifies arguments based on textual context. In several instances, the model fails to correctly identify the label when the classification depends on explicit temporal expressions such as specific years (e.g., "2017") or abbreviated quarters (e.g., "Q4"). Additionally, when an argument contains multiple temporal references such as both short-temporal expressions (e.g., "this quarter," "last two quarters") and long-term ones (e.g., "last year", "the prior year"), the model tends to categorize the argument as a short past reference.

Adding knowledge-based and TF-IDF-derived temporal features to BERT did not substantially improve. In fact, these additional features often propagated the same classification patterns observed in the BERT model. The TF-IDF feature vectors, which are intended to encode temporal information, sometimes mislead the model, causing it to incorrectly predict labels that were previously classified correctly using only BERT.

In some instances, the BERT + knowledge-based model successfully predicts the correct label when explicit temporal expressions such as "Q2" are present, as the rule-based component helps identify the temporal distance effectively. However, the model tends to underperform in cases where there is no explicit temporal reference (i.e., arguments that should be labeled as 0). Unlike the vanilla BERT model, which sometimes correctly classifies such examples based on contextual cues, including temporal distance features biases the model toward predicting labels 1 or 2. As a result, while beneficial for certain scenarios, the added temporal information contributes to performance degradation in identifying arguments with no temporal references.

5.2 FinArg-2 Results

In Detection of Argument Temporal References task, we submitted our top 2 models (SCaLAR IT_ECC_1 and SCaLAR IT_ECC_2) and the results of our models are evaluated on the official test set released by FinArg-2 are shown in table 4. SCaLAR IT_ECC_1 model got the third position in this task. SCaLAR IT_ECC_1 achieved a Micro F1 score of 70.24% and Macro F1 score of 67.85%, demonstrating strong performance across different labels. This model generalized well across different classes. In contrast, SCaLAR IT_ECC_2 exhibited significantly lower performance, with a Micro F1 score of 35.71% and Macro F1 score of 32.27%. This substantial drop in performance shows that the model is overfitting, failed to generalize the unseen data. Overall, BERTForSequenceClassification model outperforms other models highlighting its capability to effectively

classify arguments. These results indicate that end-to-end fine-tuning a pre-trained BERT model effectively captures contextual information needed for argument classification.

Table 4: Performance comparison of models on official Test Set released by FinArg-2

Submission Name	Micro F1 (%)	Macro F1(%)
SCaLAR IT_ECC_1	70.24	67.85
SCaLAR IT_ECC_2	35.71	32.27

6 Conclusion and Future Work

In this research work, we experimented with various architectures for argument classification by combining BERT with other linguistic and knowledge based features. We started with a basic BERTForSequenceClassifier as our baseline model. To further enhance performance, we incorporated temporal features extracted using a knowledge-based approach and Stanza toolkit. The results showed that BERTForSequenceClassifier achieved the highest overall performance, severing as strong baseline. Adding knowledge-based features provided moderate improvements but did not surpass the baseline. The TF-IDF based temporal features extracted using Stanza contributed to achieving more class-wise balance, with best Macro F1 score. However using full BERT embeddings for both argument and temporal text negatively impacted the performance, showing that a compact feature representation is more beneficial.

For future work, we aim to explore more sophisticated temporal reasoning techniques, including mechanisms that assign greater importance to identified temporal expressions within the argument text. Instead of appending external features like TF-IDF or rule-based outputs, we propose an integration of attention-based or weighting mechanisms that emphasize temporal entities directly within the model architecture. This would enable the model to prioritize context around temporally significant cues—such as years or quarters—and make more informed predictions. Additionally, we plan to investigate further the integration of external financial knowledge bases and temporal event linking to enhance classification accuracy. One of the main limitations of this study is the size of the dataset, with only 600 training samples available. This limited data may restrict the model's ability to generalize effectively to unseen financial arguments. Future work could focus on expanding the dataset to improve model robustness and performance.

References

[1] Ahmad Alhamzeh. 2023. Financial Argument Quality Assessment in Earnings Conference Calls. In *Lecture Notes in Computer Science*. 65–81.

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

[3] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. 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)*, Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 163–169. https://doi.org/10.18653/v1/2022.finnlp-1.22

[4] Chien-Chin Chen et al. 2018. Numeral Understanding in Financial Tweets for Fine-Grained Crowd-Based Forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 136–143. https://doi.org/10.1109/wi.2018.00-97

[5] Chien-Chin Chen, Hsin-Hsi Huang, and Hsin-Hsi Chen. 2021. Evaluating the Rationales of Amateur Investors. In *Proceedings of the Web Conference 2021*. 3987–3998. https://doi.org/10.1145/3442381.3449964

[6] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2025. Overview of the NTCIR-18 FinArg-2 Task: Temporal Inference of Financial Arguments. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2019).

[8] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

[9] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. 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)*. 1–9.