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

Sai Saketh Nandam Charan Srinivas Kumar Reddy Dasari Anand Kumar Madasamy

National Institute of Technology Karnataka Surathkal, India

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 ar- chitectures 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 tempo- ral 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 incorpo- rating 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

Introduction

The SCaLAR IT team participated in the NTCIR-18 FinArg-2 task, focusing on the Detection of Argument Temporal References subtask. This task involves classifying financial arguments from Earnings Conference Calls (ECC) dataset into:

- **Label 0**: No time reference
- **Label 1**: Long past (> 6 months)
- **Label 2**: Short past ($\leq$ 6 months)

Our goal is to enhance automated financial information processing by identifying temporal cues in arguments, aiding decision-making models.

Dataset Example

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)

Dataset

The ECC dataset includes:

- 600 training records
- 150 validation records, divided equally into 75 for validation and 75 for testing (referred to as our test set)
- Common temporal entities in financial contexts, such as "last quarter," "the last two quarters," "fiscal year," "year-over-year," "last year," "a year ago," and "the first quarter".

Methodology

We developed multiple architectures to classify temporal references in financial arguments, combining rule-based and deep learning approaches:

1. **BERTForSequenceClassifier :**
   Fine-tuned BERT (bert-base-uncased) for baseline classification. Input text (concatenated claim & premise) is tokenized with BERT Tokenizer, adding [CLS] and [SEP] tokens, with padding and truncation to a fixed length. The model predicts probabilities for labels 0, 1, & 2.
2. **BERT + KnowledgeBase:**
   Integrates BERT embeddings with temporal features extracted via regular expressions. Temporal references (e.g., quarters Q1–Q4, years) are identified and assigned numerical indices (Q1: 1, Q2: 2, etc.). Temporal distance is computed as
   - No references: Label 0, distance 0
   - Quarter differences only: Label 2 if ≤ 2 quarters, else Label 1.
   - Year differences only: Label 2 if year difference is 0, else Label 1 (distance = year difference × 4).
   - Both year and quarter: Distance = (year difference × 4) + quarter difference; Label 2 if ≤ 2, else Label 1.

   The [CLS] embedding is concatenated with two features: reference presence (binary) and temporal distance.
3. **BERT + Stanza (TF-IDF) :**
   Uses Stanzas NER package to extract DATE entities, concatenated into a temporal text string. After preprocessing (stopword removal, lowercasing, normalization), TF-IDF vectorization creates a 133-dimensional feature vector, concatenated with BERTs [CLS] embedding for classification.
4. **BERT + Stanza (BERT Tokenizer) :**
   Temporal text is tokenized using BERTs subword tokenizer, producing a 768-dimensional embedding, concatenated with the arguments [CLS] embedding. This approach leverages dense embeddings for temporal cues.

A fully connected classification layer processes concatenated features to predict label probabilities, enhancing contextual and temporal understanding

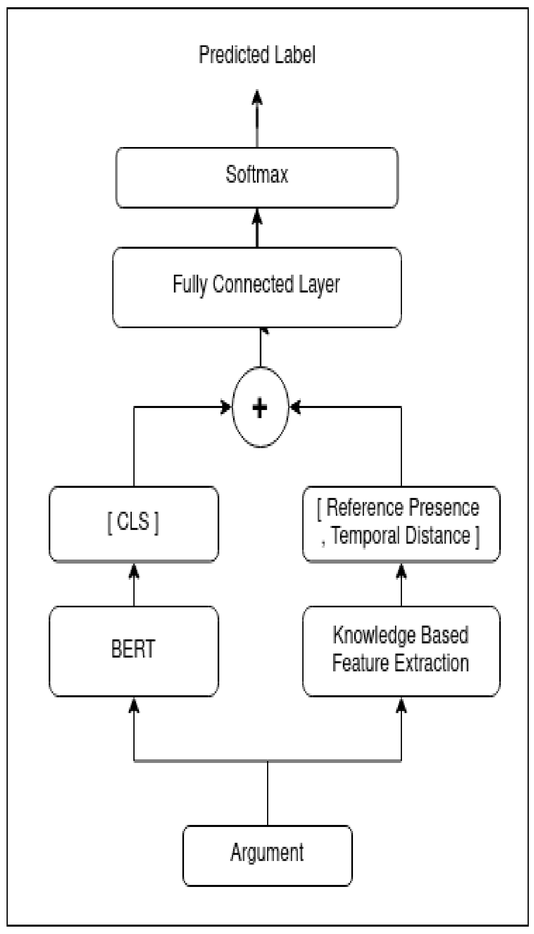

Figure 1. Architecture for the BERT + KnowledgeBase

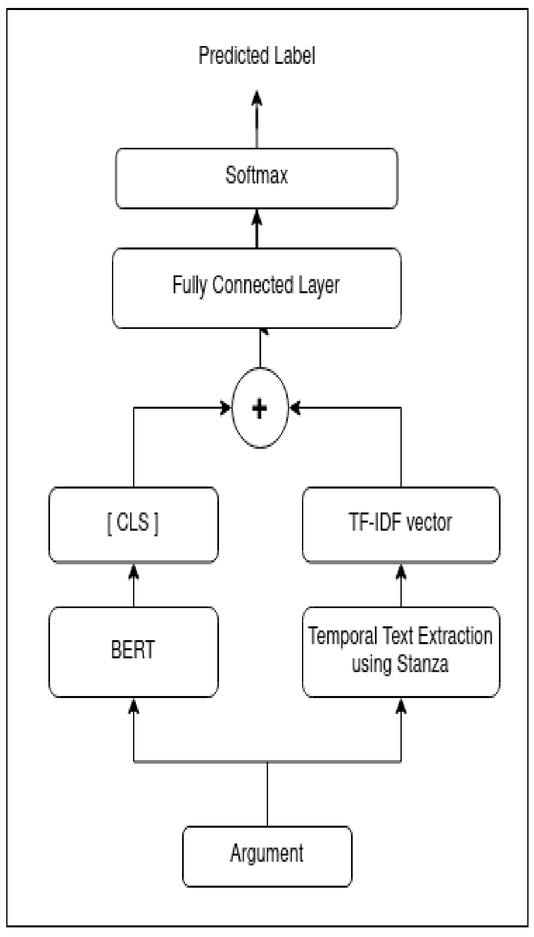

Figure 2. Model architecture for the BERT + Stanza (TF-IDF)

Results and Analysis

Table 1. Model Performance on Our Test Set

Model	Embedding Size	Micro F1 (%)	Macro F1 (%)	Weighted F1
BERTForSequenceClassifier	768	**79.00**	75.00	**79.00**
BERT + KnowledgeBase	768 + 2	76.00	75.33	77.08
BERT + Stanza (TF-IDF)	768 + 133	77.33	**75.69**	77.69
BERT + Stanza (BERT Tokenizer)	768 + 768	69.33	68.11	70.38

Table 2. Performance on Official FinArg-2 Test Set

Submission	Micro F1 (%)	Macro F1 (%)
BERTForSequenceClassifier	70.24	67.85
BERT + KnowledgeBase	35.71	32.27

- BERTForSequenceClassifier model achieved the highest Micro and Weighted F1 scores on our test set.
- BERT + Stanza TF-IDF model excelled in Macro F1, indicating balanced class performance.
- Official FinArg-2 results: Model 1 ranked 3rd, showing strong generalization, while Model 2 overfit.

Analysis Based on Experiments

- BERT relies on textual context, struggling with explicit temporal expressions (e.g., "2017", "Q4") and misclassifying arguments with mixed short/long temporal references as short past.
- Knowledge-based and TF-IDF features did not significantly improve performance and sometimes misled predictions, especially for Label 0 (no temporal reference).
- BERT + Knowledge-Based excels with explicit temporal terms (e.g., "Q2") but underperforms for Label 0 due to bias toward temporal labels

Conclusion & Future Work

- BERTForSequenceClassifier is a strong baseline, but TF-IDF-based temporal features improve class-wise balance.
- Dense BERT embeddings for temporal text reduced performance.
- **Future Work** will focus on enhancing temporal reasoning by integrating attention-based mechanisms that give higher importance to temporal expressions (e.g., years, quarters) directly within the model. This approach aims to replace external features like TF-IDF with built-in mechanisms that help the model better understand and prioritize time-related information in financial arguments.

References

[1] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning.
Stanza: A python natural language processing toolkit for many human languages.
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 58(1):101–108, 2020.

saisaketh5022@gmail.com