TMUNLPG1 AT THE NTCIR-18 FINARG-2 TASK

XUAN-YU YOU[1], LIEW DI JIE[1], WEN-CHAO YEH[2], YUNG-CHUN CHANG[1*]
[1]GRADUATE INSTITUTE OF DATA SCIENCE, TAIPEI MEDICAL UNIVERSITY
[2]INSTITUTE OF INFORMATION SYSTEMS AND APPLICATIONS, NATIONAL TSING HUA UNIVERSITY

MOTIVATION OF THE STUDY

Temporal reasoning plays a crucial role in financial argument mining, where the validity and interpretation of claims are often time-sensitive. The FinArg-2 tasks introduce challenges that require the identification of temporal references within financial arguments and the estimation of claims' validity periods. We propose two approaches leveraging domain-adapted transformer models and traditional statistical methods to address these tasks. Our goal is to investigate whether explicit temporal features enhance or hinder generalization and to assess the effectiveness of combining pretrained language models with statistical indicators for financial temporal inference.

COMPETITION & 2 TASKS

DETECTION OF ARGUMENT TEMPORAL REFERENCES (ECC SUBTASK)

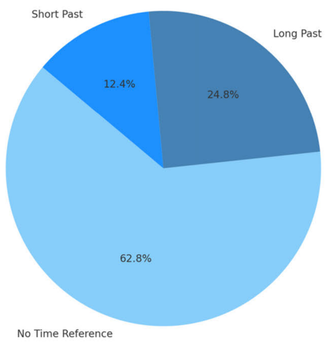
GOALS

Classify whether a financial argument from the **E**arnings **C**onference **C**all refers to a **past** specific temporal context.

LABELS

- No Time Reference (376 instances, 50.1%)
- Long Past (> 6 months ago) (216 instances, 28.8%)
- Short Past (< 6 months ago) (158 instances, 21.1%)

*Data were provided officially. 750 training data in total.



DATA SAMPLE

claim_text	premise_texts	year	quarter
trend that we see in consumer...	And there's a strong technological underpinnin...	2017	Q2

ASSESSMENT OF THE CLAIM'S VALIDITY PERIOD (SM SUBTASK)

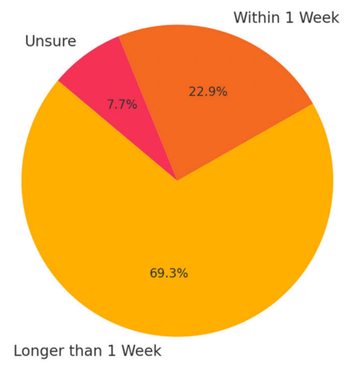
GOALS

Predict the **future** validity duration of financial claims made on **S**ocial **M**edia.

LABELS

- Longer than 1 Week (4859 instances, 69.3%)
- Within 1 Week (1607 instances, 22.9%)
- Unsure (542 instances, 7.7%)

*Data were provided officially. 7008 training data in total.



DATA SAMPLE

text
中壽可以準備賣給開發金了，除權息前應該可以完成

METHOD

Feature Engineering:
[year, quarter]⟶[time] E.g.,
[2017, Q2]⟶[This refers to year 2017, quarter 2]
- **Structure 1:** [time]+Premise+Claim
- **Structure 2:** Premise+Claim (no temporal info)

Gemini 1.5 Flash Applications:
- **Data Augmentation:** Generate synthetic samples to alleviate data imbalance challenge.
- **Prediction:** In addition to fine-tuned ModernBERT models, zero-shot and few-shot prompting approaches were explored to predict labels directly without fine-tuning.

4 Pooling Methods & Logit-Level Averaging Ensembling:
- **Average Pooling:** Averaged hidden states across selected layers.
- **Mean Pooling:** Averaged token embeddings with padding awareness.
- **CNN1D Pooling:** Applied 1D convolution followed by max pooling.
- **Weighted Pooling:** Aggregated hidden states using a learnable weight vector for each layer.

Statistical Feature Engineering:
- **Log Likelihood Ratio(LLR):** Identified keywords strongly associated with each temporal validity class by comparing term frequencies across classes. Top discriminative terms were selected and curated to avoid domain-specific noise (e.g., company names).
- **Pointwise Mutual Information (PMI):** Measures association between keywords and labels based on their co-occurrence patterns. It captures finer-grained relationships beyond term frequencies, helping the model extract semantic signals in noisy social media texts.

2 Pooling Mechanisms:
- **Average Pooling:** Uniform averaging of final-layer embeddings to create sentence-level representations.
- **Weighted Pooling:** Attention-based dynamic pooling that learned token importance weights, improving sensitivity to implicit temporal indicators

MODELS

ECC_1
- ModernBERT
- Structure 2
- Ensemble Pooling

ECC_2
- ModernBERT
- Structure 1
- Ensemble Pooling

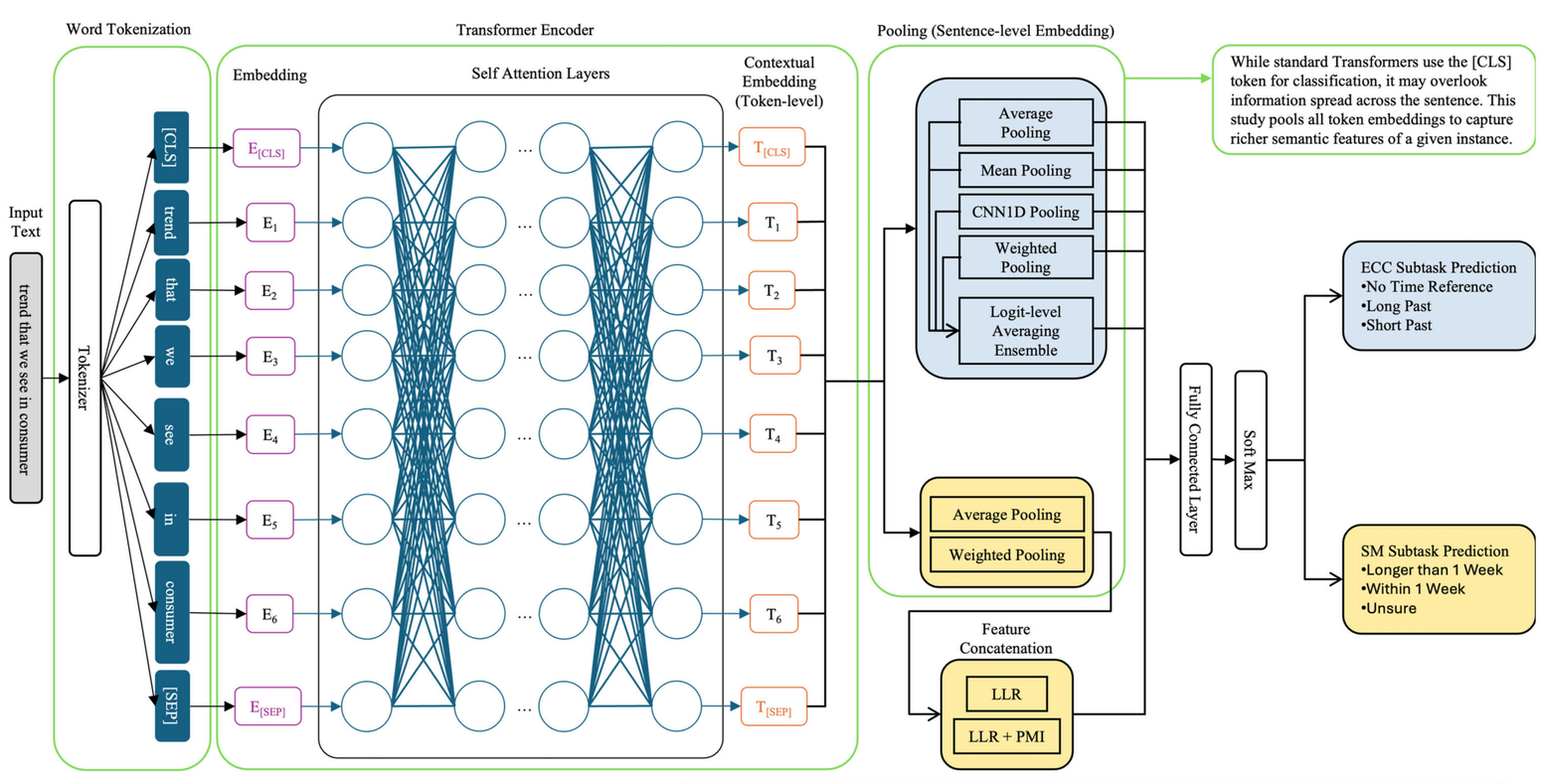
ECC_3
- Gemini 1.5 Flash
- Structure 2
- Five-shot prompting

SM_1
- FinBERT
- Weighted Pooling

SM_2
- FinBERT
- Weighted Poolig
- LLR Feature

SM_3
- FinBERT
- Weighted Poolig
- LLR + PMI Feature

MODEL ARCHITECTURE



While standard Transformers use the [CLS] token for classification, it may overlook information spread across the sentence. This study pools all token embeddings to capture richer semantic features of a given instance.

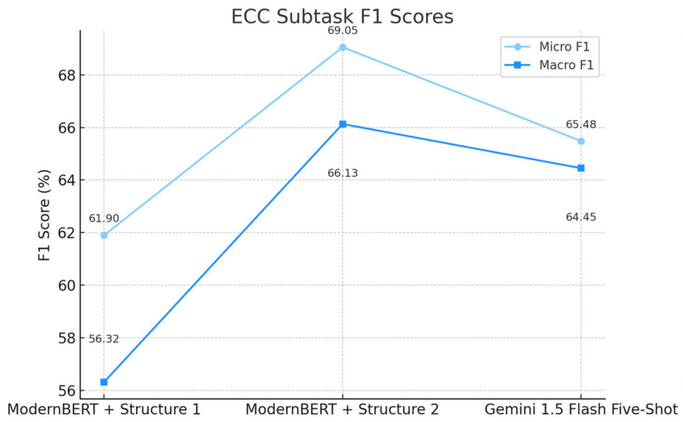
General Architecture
- After tokenization, each token is then embedded and fed to the Transformer Encoder (Self-Attention Layers).
- The output is contextual embeddings (tokens' semantic meaning within the context of the entire sentence).

ECC Subtask
- The contextual embeddings are processed through the 4 pooling method.
- Ensemble pooling: average of logits obtained from each pooling method.

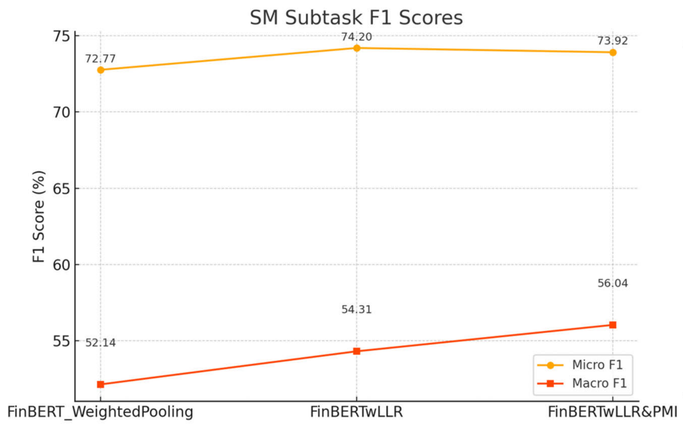
SM Subtask
- The pooled sentence representation is concatenated with LLR and PMI.

Submission Policy:
- Only 3 submissions were allowed for each task; hence, the trained models were to be tested on the unofficial test dataset. Based on the results, 3 best models were further selected for submission. The results in the line charts below are selected models' performance on official, undisclosed test data.

RESULTS & DISCUSSION



- Using multiple pooling strategies combined with ensemble averaging significantly improved performance.
- Macro F1 score increased by nearly 10% when moving from a single pooling structure to a diversified pooling structure.
- Different pooling methods capture complementary aspects of contextual information, leading to a more robust sentence representation.
- Although the Gemini 1.5 Flash five-shot model performed competitively, task-specific fine-tuning of a Transformer architecture was ultimately more effective.



- Gradual improvement was observed as handcrafted statistical features (LLR and PMI) were integrated.
- Adding LLR features enhanced performance; further incorporating PMI features led to additional gains.
- Statistical features effectively supplement deep contextual embeddings, particularly in noisy and imbalanced social media data.
- The gap between Micro and Macro F1 scores highlights challenges in accurately classifying minority classes.

The experimental results highlight two key findings:
- **In the ECC subtask:**
  - introducing multiple pooling strategies combined with ensemble averaging significantly boosted performance over a simpler single-pooling structure.
  - This demonstrates that capturing different aspects of contextual information via diverse pooling methods can effectively enhance temporal evidence identification.
- **In the SM subtask:**
  - The integration of statistical features (LLR and PMI) with the pooled contextual embeddings led to modest but consistent improvements.
  - This suggests that handcrafted linguistic knowledge, such as term association strength, complements pre-trained language models in low-resource, noisy text scenarios typical of social media data.
- These findings emphasize the importance of combining deep contextual representations with task-specific enhancements to maximize model effectiveness across different domains.

This work was supported by the National Science and Technology Council of Taiwan under grants NSTC 112-2622-E-038-001, NSTC 113-2221-E-038-019, NSTC 113-2627-M-A49-002, NSTC 113-2321-B-038-012, and NSTC 113-2321-B-038-006.