SCUNLP-3 at the NTCIR-18 FinArg-2 Task: Template-Based Prompting and Augmentation

Hong-Rui Pan, Jheng-Long Wu

Department of Data Science, Soochow University, Taiwan

NTCIR 18

Introduction

This paper presents our efforts on the NTCIR18-FinArg-2 shared task. To analyze the implicit validity of claims in Internet forum messages, we leverage a dataset where claims are annotated with predefined validity categories: valid within one week, valid for longer than one week, and unsure. Figure 1 presents an example of such a classification. In our approach, we utilize a template-based method to guide LLM-generated text, which is then combined with the original financial discussion text. This augmentation aims to enrich contextual understanding and improve the model's ability to assess validity periods more effectively.Our research explores the role of large language models (LLMs) in improving validity period prediction by leveraging contextual understanding and temporal reasoning. In Chapter 3, we discuss our approach and its implications for sentiment analysis in financial discussions. Our findings contribute to a deeper understanding of how sentiment recognition, timeliness assessment, and semantic detection interact in financial and social media texts, ultimately improving information reliability assessment in these domains.

Method

1 Upsampling

we observed that the number of instances labeled as Unsure was notably scarce. This data imbalance poses a challenge for the model in effectively distinguishing this category. To mitigate this issue, we leveraged GPT to generate a semantically equivalent text sample for each Unsure instance. By incorporating these synthetic samples into the original dataset, we effectively doubled the amount of Unsure data, thereby enhancing the model's ability to learn and generalize in recognizing this category.

2 Template-Based Prompting and Augmentation

To improve the classification of comment validity duration, we propose a structured template-based augmentation method that guides GPT to generate diverse samples with rich temporal cues and intent variations. This two-step process first extracts the core meaning related to timeliness and intent from each original comment, then uses predefined templates to generate new, consistent variants. All samples are constrained within 512 tokens to ensure compatibility with the chinese-bert-wwm-ext model. This method focus on two key challenges: identifying implicit temporal references that extend beyond the surface (Implicit Timeliness Issues) and detecting whether a comment's intent implies a specific validity period (Intent Detection Series).

The Diagram of Template-base augmentation

Determine whether the message has an implicit time limit.
Please reply in the following format:
The message <has / does not have> an implicit time limit.

Prompt for Implicit Timeliness Issues Template

Prompt for Intent Detection Series Template

Result

Result shows that upsampling significantly improves recall (from 0.528 to 0.571) and F1-score (from 0.530 to 0.577) compared to the baseline, indicating its effectiveness in addressing class imbalance and enhancing the model's ability to capture underrepresented classes. Template-based augmentation methods also lead to performance gains, particularly in accuracy and precision. Template 1 (Implicit Timeliness Issues) achieves the highest accuracy (0.781) with balanced precision and recall, while Template 2 (Intent Detection Series) yields the highest precision (0.677) but slightly lower recall. These results suggest that upsampling improves overall consistency, whereas template augmentation enriches specific textual patterns, offering complementary benefits for improving model robustness.

	acc	pre	rec	f1
Baseline	0.746	0.614	0.528	0.530
Upsampling	0.755	0.622	0.571	0.577
Template 1	0.781	0.641	0.568	0.564
Template 2	0.763	0.677	0.542	0.540

Result of accuracy, precision, f1 on Validation data

Conclusion

This study systematically evaluated the effects of upsampling and template-based prompting on validity period classification, showing that each augmentation method contributes distinct strengths. Upsampling improves recall and F1-score by addressing class imbalance, while template-based augmentation enhances the model's sensitivity to temporal and intent-related patterns, leading to gains in accuracy and precision. Our findings suggest that combining data-driven and structured augmentation approaches can provide complementary benefits. We also observed that model performance degrades with longer inputs, indicating limitations in retaining contextual information. Future work should focus on advanced architectures—such as hierarchical attention or long-context transformers—and adaptive augmentation strategies to better handle lengthy texts and further boost classification robustness.