Tokushima University

TMAK: Trustworthy Multi-modal Affective AI and Knowledge Engineering Lab

TMAK at NTCIR-18 FinArg-2 Task

Takahiro Kawamoto Xin Kang

Introduction

In FinArg-2, we extend our previous work by performing temporal reasoning on financial discussions, as understanding the temporal context of financial documents is useful for decision support. We use the same resources developed in FinArg-1, where we analyzed financial documents and proposed a method that integrates discussion mining with sentiment analysis.

Each instance in the dataset consists of the following elements: “claim_text,” “premise_texts,” “year,” “quarter,” and “label.” The “label” indicates the type of temporal reference (0: no time reference, 1: long past, 2: short past), which we used as the target for classification. Specifically, label 1 represents a temporal reference to a point more than half a year ago, while label 2 represents a reference to this quarter or up to two previous quarters.

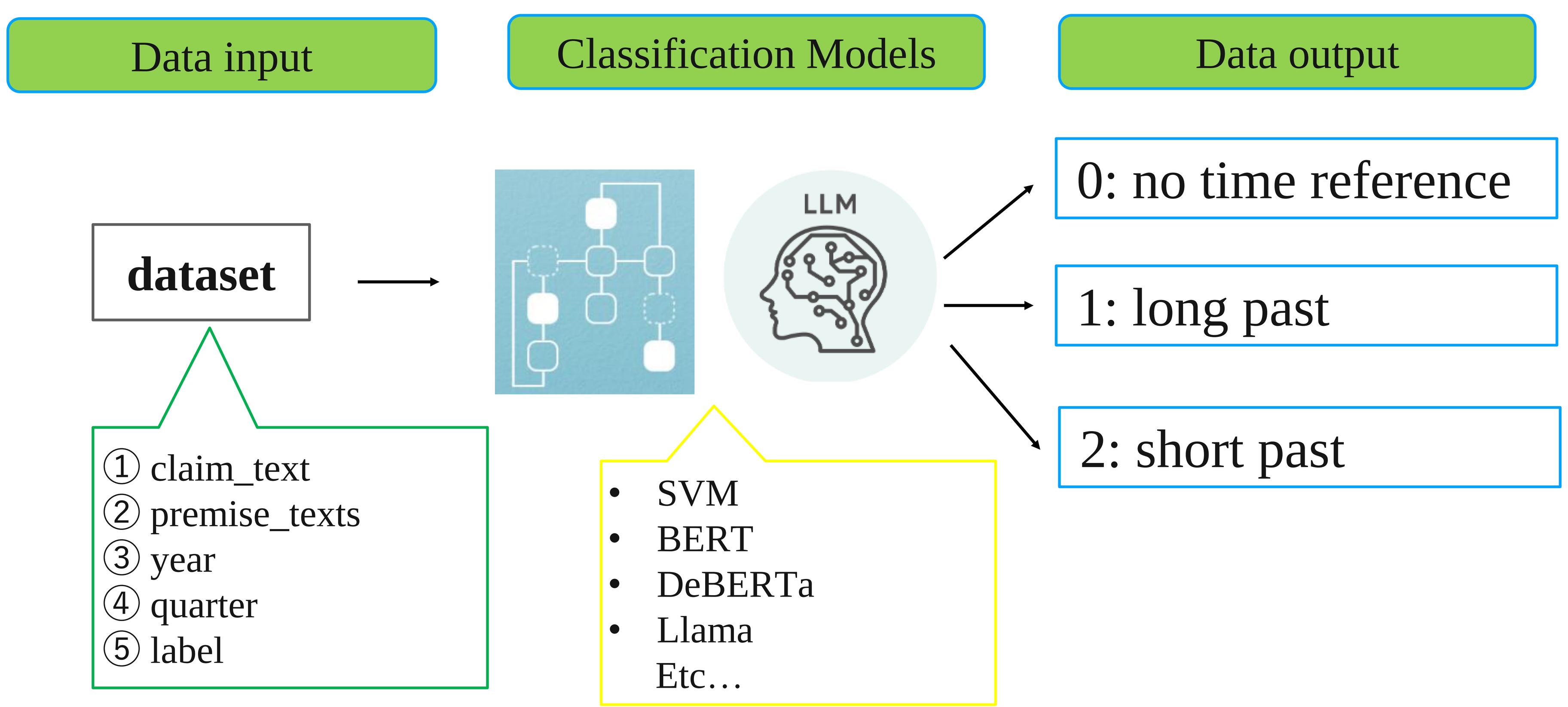

Figure 1: Flowchart of the Proposed Method

Methods

We built temporal text classifiers using two approaches: conventional machine learning and Large Language Models (LLMs). We built temporal sentence classifiers with both approaches and evaluated their performance based on classification accuracy.

Conventional machine learning models:

In this study, we used two conventional machine learning methods: Logistic Regression model and Support Vector Machines (SVM) model.

We conducted text classification using these two models with TF-IDF features.

SVM is a classification method that sets boundaries that maximize the distance between the boundaries that serve as the classification criteria for classes and each piece of data. Logistic Regression is a data analysis technique that uses mathematics to find the relationship between two data factors. Then, this relationship is used to predict the value of one factor based on the other.

The textual data is treated as mathematical data by vectorizing it with TF-IDF.

Large-scale language model:

In this study, we performed a comparative analysis using BERT and its derived models, as well as the Llama model.

There are various types of LLMs, and we adjusted the parameters and data format to suit them. It was very interesting to learn about model selection and parameter adjustment, as the output results are limited depending on the results of the learning.

Results

The performance of each model was evaluated in terms of classification Accuracy, Micro-F1 score, and Macro-F1 score. The results are shown in Tables 1, 2, and 3, respectively.

The models based on LLMs showed better performance than the models based on conventional machine learning.

Table 1: Model Results

Model	Epochs	Learning Rate	Accuracy
Logistic Regression	-	-	0.6800
SVM	-	-	0.6700
BERT	3	5e-5	0.7500
RoBERTa	6	5e-5	0.7670
DeBERTa	8	3e-5	**0.7800**
Llama	5	3e-4	**0.7800**

Table 2: Comparison Based on Different Parameters

Model	Epochs	Learning Rate	Accuracy
Llama	6	1e-4	0.5733
Llama	5	3e-4	**0.7800**

Table 3: Comparison of Micro-F1 and Macro-F1 scores of Conventional Machine Learning and LLMs Method

Model	Micro-F1	Macro-F1
Logistic Regression	0.6071	0.5275
BERT	**0.6310**	**0.5867**

Conclusion and Future Work

In this study, we considered various methods to devise a model to classify the temporal relationships between sentences and found that extremely high performance can be achieved by using LLMs. Classification models using LLMs performed better than classification models using conventional machine learning. We also found that LLMs can show different results even for the same model by using appropriate parameters such as the learning rate and the number of epochs.

In future work, we will continue test models to find better performing models and optimal parameters. We will run them with lower parameters first, and for models that perform similarly to others, we will experiment with higher parameters. We will also look further into methods that we have not yet used, such as relatively new LLMs such as DeepSeek, and then validate them with what is available.