TicTac: Time-aware Supervised Fine-tuning for Automatic Text Dating

Han Ren[1,2] **Minna Peng**[3]

[1]Laboratory of Language and Artificial Intelligence, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies
[2]Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies
[3]School of Information Science and Technology, Guangdong University of Foreign Studies
hanren@gdufs.edu.cn, pengminna@mail.gdufs.edu.cn

Abstract

Pre-trained language models have achieved success in many natural language processing tasks, whereas they are trapped by the time-agnostic setting, impacting the performance in automatic text dating. This paper introduces TicTac, a supervised fine-tuning model for automatic text dating. Unlike the existing models that always ignore the temporal relatedness of documents, TicTac has the ability to learn temporal semantic information, which is helpful for capturing the temporal implications over long-time span corpora. As a fine-tuning framework, TicTac employs a contrastive learning-based approach to model two types of temporal relations of diachronic documents. TicTac also adopts a metric learning approach, and the temporal distance between a historical text and its category label is estimated, which is beneficial to learning temporal semantic information on texts with temporal ordering. Experiments on two diachronic corpora show that our model effectively captures the temporal semantic information and outperforms state-of-the-art baselines.[1]

1 Introduction

The temporal dimension of texts is essential to many time-sensitive tasks, such as question answering (Shang et al., 2022; Son and Oh, 2023), natural language inference (Vashishtha et al., 2020; Liu et al., 2021), event detection (Hettiarachchi et al., 2022), and text generation (Cao and Wang, 2022; Pratapa et al., 2023). Time-based reasoning in these tasks benefits from timestamps or temporal metadata of documents, which, however, are not always available (Chambers, 2012). One way to tackle this issue is automatic text dating (ATD), which refers to the task of identifying when a document was written according to its content (Dalli, 2006).

A straightforward way of ATD is to find temporal clues within texts. For example, an article with a sentence *the novel was published in* 2021 suggests that it was written no earlier than 2021. However, it is more challenging when explicit temporal mentions do not appear. To address it, research on ATD focuses on representation learning, trying to find temporal clues from diachronic changes in language. This work has a wide range of application scenarios in the digital humanities (Baledent et al., 2020). One example is to identify the dates of historical documents in digital libraries where temporal metadata are missing, which is also known as historical text dating (Popescu and Strapparava, 2015; Boldsen and Wahlberg, 2021).

Current studies on historical text dating usually adopt pre-trained language models (PLMs) to representation learning (Tian and Kübler, 2021; Li et al., 2023). Although such models have achieved success in many natural language processing tasks, they are trapped by the time-agnostic setting, impacting the performance in time-sensitive downstream tasks, including text dating (Röttger and Pierrehumbert, 2021; Su et al., 2022). To address it, recent efforts on diachronic text modeling either train large language models on diachronic corpora (Wang et al., 2023, 2024), trying to improve their performances on time-sensitive tasks, or employ sophisticated learning models to capture lexical semantic changes over time (Ren et al., 2023; Wei et al., 2025). However, the prominent limitation of the former is the limited task data, high training costs, and the risk of catastrophic forgetting, while the latter ignores the temporal relatedness of diachronic documents, which may be helpful for capturing temporal semantic variations. Besides, temporal categories in most studies are viewed as independent from each other rather than sequentially related, impacting the ability of temporal-based text modeling.

In this paper, we propose TicTac, a supervised

[1]Our code is available at: https://github.com/gdufslec/Tictac.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21906–21918

fine-tuning model for ATD. This work attempts to tackle three aforementioned issues in current studies: 1) how to develop a method to learn temporal semantic information at low training costs; 2) how to capture language evolution features based on temporal-related documents and; 3) how to model the sequential relatedness of temporal categories for ATD. To this end, our model firstly adopts a supervised fine-tuning framework to learn from diachronic documents, which avoids high computational costs caused by training from scratch and offers boarder applicability to downstream tasks as well. Secondly, we propose a contrastive learning-based diachronic document modeling approach, trying to capture temporal implications in temporal-related documents. For example, two documents discussing different technological advances of spinning machines and the Internet suggest that they may be written in different eras. Specifically, we define two types of temporal relations: one is relative temporal relations between documents, and the other is absolute temporal relations between documents and temporal category labels. This idea contributes to a better understanding of the relationship between time and language evolution. Thirdly, we treat ATD as an ordinal classification task and propose a corresponding approach to improve the performances by appropriately correcting prediction errors of different degrees. In this approach, the ordinal relationship between temporal categories is considered, which helps to promote evaluation accuracy and thus benefits the modeling of diachronic documents with temporal ordering. Our contributions are summarized as follows:

- We propose TicTac, a fine-tuning model for ATD, to learn temporal semantic information from diachronic documents in a supervised way.
- We propose a contrastive learning-based time-ordered document modeling approach from two temporal perspectives, trying to capture temporal implications in temporal-related documents.
- We propose an ordinal classification approach for the ATD task, where the ordering of temporal categories is considered for the classification task.
- Experiments on two diachronic datasets show that our model effectively captures the temporal semantic information and outperforms state-of-the-art baselines.

2 Related Work

2.1 Automatic Text Dating

Early studies on ATD either rely on manually extracted temporal features to identify explicit temporal expressions (Dalli, 2006; Kanhabua and Nørvåg, 2008; Niculae et al., 2014) or employ traditional machine learning methods to explore statistical features for temporal classification models (Garcia-Fernandez et al., 2011; Ciobanu et al., 2013; Boldsen and Wahlberg, 2021), whereas most of them suffer from low generalization ability and are insufficient to capture implicit temporal clues within texts.

Deep learning methods help to learn implicit temporal features from texts, improving the performance in the ATD task (Ray et al., 2018; Yu and Huangfu, 2019a). Large pre-trained language models such as SentenceBERT (Massidda, 2020; Tian and Kübler, 2021) and RoBERTa (Li et al., 2023) significantly promote ATD by leveraging on training with large-scale data. However, these models often ignore the dynamic temporal attributes of word meanings, limiting the ability to detect semantic distinctions between documents of different periods. To address it, Ren et al. (2023) proposed a time-aware language model (TALM), aiming to capture implicit temporal information to acquire better word representations. However, they overlook the temporal relatedness between diachronic documents, which may help to capture temporal semantic variations.

2.2 Ordinal Classification

Ordinal Classification (OC) aims to categorize data according to a predefined sequence of ordered or ranked labels. Unlike general classification tasks, which treat categories as independent ones, OC considers the ordering between categories (Amigo et al., 2020). Methods on OC tasks often model the relationships between categories, including their order and distance. They usually employ metric losses to represent the distances between categories. For example, Hou et al. (2016) proposed Squared Earth Mover's Distance-based Loss (EMD), which quantified inter-category differences and was particularly effective for tasks with ordinal relations. Torre et al. (2018) introduced Weighted Kappa Loss (WKL), a loss function grounded in

the weighted kappa coefficient, for multi-class classification with ordered labels. Diaz and Marathe (2019) developed a soft-label (SOFT) approach, where labels were represented as probability distributions, enabling the model to learn the ordinal relations without changing the network architecture when combined with standard classification losses (e.g., cross-entropy). Castagnos et al. (2022) introduced Ordinal Log Loss (OLL), a method to address the limitations of traditional cross-entropy loss in ordered text classification tasks. Kasa et al. (2024) proposed Multi-task log loss function (MLL), which integrated OLL with cross-entropy (CE), optimizing a hybrid loss function to more effectively capture the ordinal information across categories. Given that temporal categories in the ATD task have the characteristics of ordering, it is more suitable to model temporal ordering of diachronic documents for a better ATD performance.

3 Methodology

3.1 Task Formulation

ATD aims to assign a temporal label to each data. Given a document $d = \{w_1, w_2, \ldots, w_l\}$, where w_i denotes the i-th word and l represents the document length, the task is to predict a temporal label t from a predefined set of time categories $T = \{t_1, t_2, \ldots, t_n\}$, each of which is a time period. Such task can be viewed as a classification problem, and the objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{1}$$

where $\hat{y}_i$ represents the predicted probability for the i-th document, and y_i the true label of it. By minimizing the cross-entropy loss function $\mathcal{L}_{\text{CE}}$, the model is trained to identify temporal semantics within texts, thereby predicting temporal categories to each document.

3.2 Overall Architecture

This section provides a detailed description of the overall architecture of the proposed method TicTac, as illustrated in Figure 1. Our model fine-tunes PLMs on the ATD task. Specifically, it learns the representations by capturing temporal information via two parts: time-aware contrastive learning and time-aware ordinal classification.

3.3 Time-aware Contrastive Learning

Contrastive learning in our model aims to capture temporal implications within diachronic documents. We define a fine-tuning framework for contrastive learning, and the input documents are first tokenized and processed by a PLM encoder to get contextualized representations:

$$H = \text{BERT}(X) = \{h_{[\text{CLS}]}, h_1, h_2, \ldots, h_l\}, \tag{2}$$

where $h_{[\text{CLS}]}$ serves as a global representation of an input document.

The objective of the fine-tuning process is to model similarities and differences among texts, thereby enhancing the model's capacity to capture temporal semantic relations within them. To this end, we propose a supervised contrastive learning process (Gunel et al., 2020) to learn the semantic association of diachronic documents. Such association can be decomposed as two types of relations: relative temporal relation and absolute temporal relation. The former represents local relations between documents, while the latter denotes global ones between documents and temporal categories.

3.3.1 Contrastive Learning via Relative Temporal Relation

To capture relative temporal relations between texts from different diachronic contexts, we introduce a supervised contrastive learning approach that clusters samples(i.e., documents) with the same temporal categories while separating those from different categories. This approach enables the model to learn local semantic relations within documents.

Specifically, for each document i, we define a positive sample set $\mathcal{P}_i$, which consists of documents from the same temporal category, and a candidate sample set $\mathcal{A}_i$, which includes all the other samples in the batch, excluding i. The learning objective is formulated as:

$$\mathcal{L}_{CLR} = \frac{1}{N} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} -\log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in \mathcal{A}_i} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}, \tag{3}$$

where N is the number of documents in the batch, $\mathcal{I}$ is the index set of all batch documents, and τ is the temperature parameter that adjusts the sensitivity of the contrastive loss. The terms $\boldsymbol{z}_i$ and $\boldsymbol{z}_p$ represent the CLS embeddings of the anchor sample i and its positive sample p, respectively.

By optimizing this objective, the model captures similarities within documents of the same temporal categories and differences within documents from

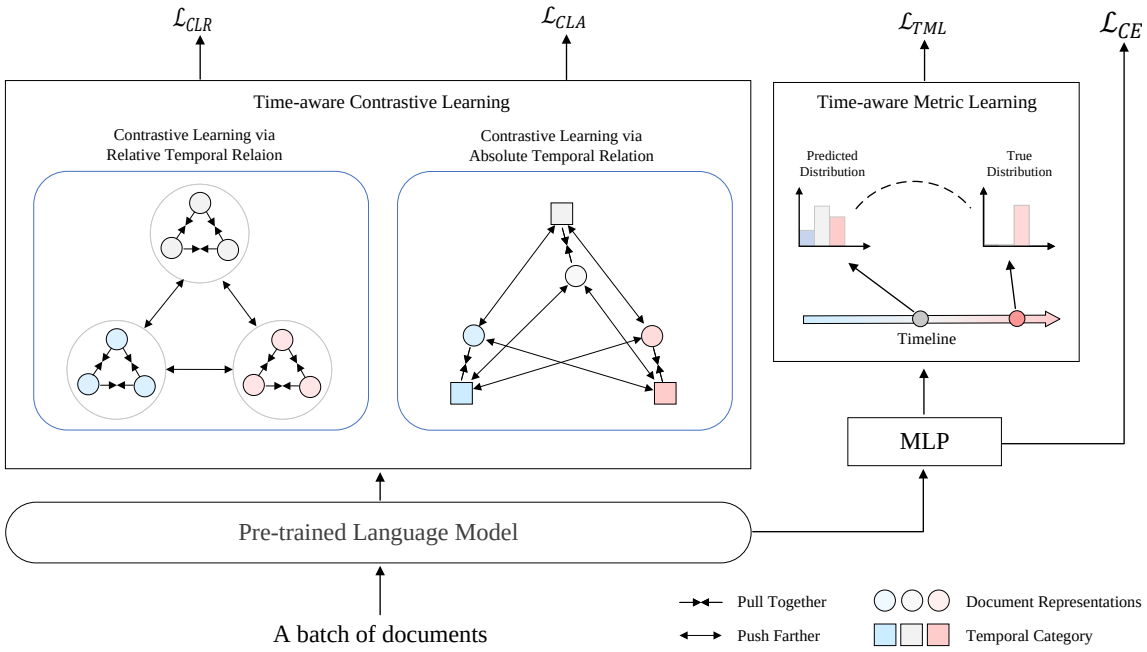

Figure 1: An overall architecture of TicTac.

different ones to gain representations with relative temporal semantics.

3.3.2 Contrastive Learning via Absolute Temporal Relation

The temporal relationship between documents and temporal categories regarded as the absolute temporal relation provides an alternative perspective to the relative temporal relation for illustrating the semantic association between diachronic data. To this end, we firstly propose to learn a label representation for each time category. Then, we maximize the distance between documents and their corresponding category labels while minimizing the distance between documents and the categories they do not belong to. By this approach, documents are encouraged to cluster around their respective temporal category centers.

To learn label embeddings, we introduce a learnable weight matrix W_{label}, which stores the embedding vectors corresponding to each label. For a given temporal category t, the label embedding $\boldsymbol{z}_l$ is obtained by a lookup operation on the matrix W_{label}, retrieving the appropriate embedding for that category. It guarantees that each temporal label is uniquely represented in the high-dimensional embedding space.

Formally, the label embedding for each temporal category t is defined as:

$$\boldsymbol{z}_l = W_{\text{label}}[t], \tag{4}$$

where W_{label} denotes the learnable matrix of label embeddings, and t represents the index of the temporal category. This lookup operation retrieves the corresponding label embedding $\boldsymbol{z}_l$, which is subsequently utilized within the contrastive learning framework. The label-centered contrastive learning objective is thus formulated as:

$$\mathcal{L}_{CLA} = \frac{1}{N} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} -\log \frac{\exp\left(\boldsymbol{z}_l \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in \mathcal{A}_i} \exp\left(\boldsymbol{z}_l \cdot \boldsymbol{z}_a / \tau\right)}. \tag{5}$$

3.4 Time-aware Ordinal Classification for ATD

Cross-entropy is a common loss function that focuses on achieving an exact match between predicted and true labels. However, it may neglect the ordinality of temporal categories. Consider a document actually written between 1800-1820: a misclassification to 1820-1840 should incur a smaller penalty than a more distant error (e.g., 1840-1860), as the former has a better temporal approximation.

To overcome the limitation, we propose to adopt Earth Mover's Distance (EMD) as a distribution-based evaluation metric. By utilizing the cumulative distribution function (CDF), EMD more accurately measures the divergence between predicted

and ground-truth temporal distributions. This approach provides a more sophisticated characterization of ordinal relationships among temporal categories compared to conventional metrics.

Specifically, we define the predicted CDF, $\text{CDF}_{\text{pred}}(k, i)$, as the cumulative sum of the predicted probabilities for sample i up to class k. The CDF effectively captures the distribution of temporal probabilities across the classes, as shown in the equation:

$$\text{CDF}_{\text{pred}}(k, i) = \sum_{j=1}^{k} \hat{y}_{i,j}, \quad k = 1, \ldots, |T|, \quad (6)$$

where $\hat{y}_{i,j}$ is the predicted probability that sample i belongs to class j, and $|T|$ represents the total number of temporal categories. In contrast, the true CDF, $\text{CDF}_{\text{true}}(k, i)$, is determined by the ground-truth temporal label, y_i, and is defined as:

$$\text{CDF}_{\text{true}}(k, i) = \begin{cases} 0, & \text{if } k \leq y_i \\ 1, & \text{if } k > y_i, \end{cases} \quad (7)$$

where $y_i \in \{1, \ldots, |T|\}$ is the true temporal label for sample i.

To minimize the difference between the predicted and true temporal distributions, we compute the EMD loss as follows:

$$\mathcal{L}_{TOC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|T|} \left(\text{CDF}_{\text{pred}}(k, i) - \text{CDF}_{\text{true}}(k, i) \right)^2. \quad (8)$$

3.5 Training Objective

The overall optimization objective $\mathcal{L}$ is a combination of multiple loss functions to concern different aspects of the ATD task mentioned above. Specifically, it is formalized as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{CLR} + \mathcal{L}_{CLA} + \mathcal{L}_{TOC}. \quad (9)$$

Through the joint optimization, our model simultaneously achieves three key objectives: (1) accurate temporal label classification, (2) preservation of ordinal relationships between time periods, and (3) improved quality of learned temporal embeddings.

4 Experiments

4.1 Datasets and Metrics

We evaluate the performance of our model on two diachronic datasets: Twenty-Four Histories Corpus (Zinin and Xu, 2020) and Royal Society Corpus (Kermes et al., 2016). The Twenty-Four Histories Corpus is a Chinese historical corpus spanning from 2500 B.C. to 1600 A.D., consisting of 2,647 volumes covering a wide range of topics related to ancient China. Each volume contains approximately 8,000 characters, resulting in a total of around 40 million characters. Since some of these texts are compiled by later authors rather than direct accounts from the periods they describe, timestamps are assigned based on the documents' publishing dates. The corpus has been divided into discrete historical periods, with each time category corresponding to a fixed time span determined by historical conventions. The Royal Society Corpus is an English-language collection with 9,779 documents covering the period from 1660 to 1880. However, the data collection does not contain predefined periodization. Following Ren et al. (2023), we divide it into 11 categories, each covering a distinct time span of 20 years.

In our experiment, texts from both datasets are segmented into text blocks of approximately 420 characters. Documents with insufficient text contents are removed to ensure the quality of the training data. Following Ren et al. (2023), we split each dataset into training, validation, and test sets in an 8:1:1 ratio. For evaluation, we adopt five metrics: precision (P), recall (R), F1 score, classification accuracy (C-acc), and adjacent accuracy (A-acc).

4.2 Baselines

The baseline models employed in the experiments include non-pretrained models (LSTM, TALM), pretrained models (BERT, RoBERTa, SBERT), ordinal classification models (WKL, OLL), and large language models (Qwen2.5, Baichuan2, GPT-4o-mini). The selection of pretrained models follows the same configuration as Ren et al. (2023) for a fair and consistent comparison. The prompts used for large language models are detailed in Table 4 in Appendix A.

We gives a brief overview of the baseline models employed in this study a follows: (1) **LSTM** (Yu and Huangfu, 2019b): a widely used recurrent neural network designed to effectively handle time-series data and sequential dependencies, making it suitable for ATD task. (2) **TALM** (Ren et al., 2023): incorporates a temporal alignment and adaptation module, optimizing the model's performance in ATD task by better capturing temporal dynamics. (3) **BERT** (Devlin, 2018): a Transformer-based

pre-trained language model that has become a standard in natural language processing, demonstrating strong performance across a range of tasks. (4) **SBERT** (Tian and Kübler, 2021): an extension of BERT that utilizes a twin network architecture to optimize sentence-level embeddings, significantly improving performance in period classification of Chinese historical texts. (5) **RoBERTa** (Li et al., 2023): an enhanced version of BERT, trained on a larger dataset with extended training time, resulting in improved performance for chronological classification of ancient Chinese texts. (6) **WKL** (Torre et al., 2018): an ordinal classification approach that leverages KL divergence to measure the differences among temporal categories, with BERT serving as the underlying model architecture in this study. (7) **OLL** (Castagnos et al., 2022): an ordinal classification approach that enhances traditional classification loss functions by explicitly accounting for the ordinal relationships between categories by using BERT as the model architecture in the experiment. (8) **Qwen2.5** (Yang et al., 2024): a powerful generative language model, capable of addressing complex natural language reasoning tasks, particularly effective for analyzing historical texts, with a 7B parameter model used in this study. (9) **Baichuan2** (Yang et al., 2023): trained on a high-quality dataset consisting of 2.6 trillion tokens, with a 7B parameter model selected for this study. (10) **GPT-4o-mini** (OpenAI, 2024): a compact variant of the GPT-4 architecture, designed for efficient text comprehension and generation, particularly suited for classification tasks requiring detailed text representations.

4.3 Implementation Details

During the training stage, both the dimensions of the BERT model and the label embedding $\boldsymbol{z}_l$ are set to 768-dimensional vectors. The dropout rate is 0.5, the learning rate is 1×10^{-5}, and the AdamW optimizer is employed. The batch size for the training, validation, and test sets is 32. Early stopping is applied with a patience of 5 epochs. In both contrastive learning modules, τ is set to 0.1. The training is performed on a single RTX 4090 GPU with 24 GB of memory, and the CPU configuration includes 10 vCPUs.

4.4 Overall Results

Table 1 shows the performance of the proposed model TicTac and baseline models on two diachronic datasets, namely the Twenty-Four Histories Corpus and the Royal Society Corpus. TicTac outperforms baseline models in almost all evaluation metrics on both datasets, illustrating a strong benchmark for the ATD task.

The results also reveal some findings. First of all, non-pretrained models including LSTM and TALM do not perform well. Specifically, two models achieve an F1 score of 72.59% and 73.91% on the Twenty-Four Histories Corpus, and an F1 score of 52.91% and 55.76% on the Royal Society Corpus, respectively. Such performances are significantly lower than TicTac, showing the insufficient ability for ATD without work on the time-aware fine-tuning paradigm.

Secondly, PLMs illustrate superior performances in most cases. Specifically, RoBERTa achieves the highest F1 score of 87.94%, closely followed by SBERT (87.56%) and BERT (86.32%) on the Twenty-Four Histories Corpus. The superior performance of RoBERTa can be attributed to its extensive pre-training and optimized training strategy, enabling it to capture nuanced contextual information in historical texts more effectively. On the other hand, PLMs do not achieve as good performance increase on the Royal Society Corpus as the Twenty-Four Histories Corpus, compared to non-pretrained models. It is mainly because ancient English words occur in the Royal Society Corpus while they do not appear in the modern English data employed for pre-training by PLMs, thereby the impact on PLM performances may be derived from unknown words.

Moreover, ordinal classification models including WKL and OLL exhibit unsatisfying performances, particularly with respect to C-accuracy and A-accuracy. Specifically, WKL records a relatively low F1 score of 53.35% and a C-accuracy of 63.74% on the Twenty-Four Histories Corpus. Although OLL shows slight improvements, it still underperforms in comparison with PLMs. It indicates that it may be necessary to introduce appropriate learning objectives for the ATD task.

Finally, sophisticated large language models show very limited ability on the ATD task. Despite their strong generative capabilities, they face significant challenges in the ATD task: Qwen2.5 attains an F1 score of only 8.46% and a C-accuracy of 7.81%, while other LLMs similarly underperform. It shows the limitation of large language models in tasks that require the ability of deep understanding and analysis, emphasizing the advantage of task-specific architectures like TicTac in this domain.

Method	Twenty-Four Histories Corpus					Royal Society Corpus				
	P	R	F1	C-acc	A-acc	P	R	F1	C-acc	A-acc
LSTM (Yu and Huangfu, 2019b)	74.13	71.70	72.59	76.70	87.92	55.17	54.48	52.91	58.60	87.52
TALM (Ren et al., 2023)	77.50	71.81	73.91	78.64	90.36	57.04	55.78	55.76	59.28	85.48
BERT (Devlin, 2018)	87.26	85.76	86.32	88.57	93.03	59.87	57.95	58.14	62.49	89.17
SBERT (Tian and Kübler, 2021)	87.72	87.74	87.56	89.41	94.23	61.24	58.34	58.68	62.72	90.19
RoBERTa (Li et al., 2023)	88.11	**88.06**	87.94	89.39	94.08	60.07	60.15	59.96	63.34	89.02
WKL (Torre et al., 2018)	51.50	61.43	53.35	63.74	92.59	44.45	46.86	44.40	51.58	89.40
OLL (Castagnos et al., 2022)	86.51	85.22	85.73	87.95	93.65	59.91	59.22	59.37	61.37	90.63
Qwen2.5 (Yang et al., 2024)	9.18	11.62	8.46	7.81	58.83	30.07	14.87	12.21	13.55	30.91
Baichuan2 (Yang et al., 2023)	4.09	5.22	3.74	3.90	52.05	31.71	15.15	16.74	16.13	37.47
GPT-4o-mini (OpenAI, 2024)	9.94	10.82	6.81	6.79	48.62	26.22	18.95	16.14	17.16	35.31
TicTac (ours)	**89.17**	87.67	**88.36**	**90.18**	**95.01**	**64.14**	**62.43**	**62.60**	**67.66**	**91.95**

Table 1: Performance comparison between TicTac (our model) and baseline models on the Twenty-Four Histories corpus and the Royal Society Corpus. The baseline models include non-pretrained models (LSTM, TALM), PLMs (BERT, SBERT, RoBERTa), ordinal classification models (WKL, OLL), and large language models (Qwen2.5, Baichuan2, GPT-4o-mini). Evaluation metrics include precision (P), recall (R), F1 score, C-accuracy (C-acc), and A-accuracy (A-acc).

Dataset	Model	P	R	F1
Twenty-Four Histories Corpus	TicTac (ours)	**89.17**	87.67	**88.36**
	w/o CLR	88.22	**87.89**	87.98
	w/o CLA	88.32	86.86	87.37
	w/o TOC	88.31	86.35	87.19
Royal Society Corpus	TicTac (ours)	**64.14**	**62.43**	**62.60**
	w/o CLR	63.90	61.37	61.93
	w/o CLA	62.65	61.43	61.53
	w/o TOC	62.89	61.85	61.88

Table 2: The results of the ablation study on the Twenty-Four Histories Corpus and Royal Society Corpus. CLR refers to the Contrastive Learning via Relative Temporal Relation module, CLA refers to the Contrastive Learning via Absolute Temporal Relation module, and TOC refers to the Time-aware Ordinal Classification for ATD module.

In summary, TicTac outperforms all state-of-the-art baselines, illustrating its effectiveness in the ATD task. It represents the significance of modeling temporal relationships for temporal texts, which is a critical issue for the ATD task.

4.5 Ablation Study

Table 2 shows the results of the ablation study on the two datasets, revealing the contribution of each module in our proposed model. It clearly demonstrates that each module contributes positively to the model's overall performance. Specifically, removing the CLR module leads to a slight performance decrease in both datasets, with the F1 score on the Twenty-Four Histories Corpus dropping from 88.36% to 87.98%. Likewise, removing the CLA and the TOC modules results in further performance declines, with F1 scores decreasing to 87.37% and 87.19%, respectively. Similar performance results are observed on the Royal Society Corpus, where removing each module causes a reduction in precision and F1 score. On the other hand, the observed performance degradation is more gentle. For example, the w/o CLR variant experiences a drop in F1 from 62.60% to 61.93%, and the removal of CLA and TOC similarly results in modest performance decreases. It indicates that the Royal Society Corpus presents a relatively more challenging benchmark dataset for the ATD task than the other one, partly because of the shorter time span for each category in it.

Overall, the results highlight the significant contribution of all three modules in effectively modeling temporal relationships within the diachronic documents. The CLR and the CLA module enhance the model's ability to capture temporal implications by leveraging relative and absolute temporal relations, respectively, while the TOC module strengthens the model's ability to learn ordinal relations within diachronic documents. Since the performance decreases when these modules are removed, it indicates the critical role these components play in achieving the superior results.

4.6 Comparison of Ordinal Classification Losses

Table 3 compares the performances of models by leveraging on different ordinal classification loss functions, based on TicTac. It can be seen from the table that, TicTac achieves the highest F1 scores, with 88.36% on the Twenty-Four Histories Corpus and 62.60% on the Royal Society Corpus, showing

Model	Twenty-Four Histories Corpus			Royal Society Corpus		
	P	R	F1	P	R	F1
EMD (TicTac)	**89.17**	87.67	**88.36**	**64.14**	**62.43**	**62.60**
OLL (Castagnos et al., 2022)	86.54	86.08	86.22	62.47	60.22	60.71
WKL (Torre et al., 2018)	87.64	87.31	87.38	62.66	61.22	61.33
SOFT (Diaz and Marathe, 2019)	88.25	**88.31**	88.21	62.94	61.38	61.84

Table 3: Comparison of different ordinal classification losses on the Twenty-Four Histories Corpus and the Royal Society Corpus. For the OLL method, α is set to 1.5, and for SOFT method, β is set to 3, as these parameter settings achieved the best performance on the Amazon reviews dataset (Castagnos et al., 2022).

its capacity in capturing complex temporal relations by employing the Earth Mover's Distance (EMD) loss for ordinal classification.

OLL underperforms on both two datasets, achieving a 2.14% and a 1.89% F1 score lower than TicTac, respectively. Although WKL performs better than OLL, it still has a 0.98% and a 1.27% F1 score lower than TicTac, respectively. Such results suggest the insufficient ability of these two models in modeling temporal ordinality in the ATD task.

SOFT employs soft labels to represent ordinal relationships, yielding a better performance in comparison with OLL and WKL. Specifically, it achieves an F1 score of 88.21% and 61.84% on the Twenty-Four Histories and Royal Society corpora, respectively. However, SOFT essentially forces models to make a regularization process rather than to learn temporal semantic relations between diachronic documents. Hence it cannot enable models to have the capacity of capturing temporal implications on temporal data.

4.7 Visualization Analysis

We display t-SNE visualizations of the CLS embeddings learned by our proposed model and the BERT model on both two datasets, shown in Figures 2 and 3 in the Appendix A.

We can find two points from the figures. First of all, there are more local clusters of document representations on the Twenty-Four Histories Corpus than the Royal Society Corpus, no matter what model the experiment conducts, suggesting that the clustering performances of all models on the Twenty-Four Histories Corpus are better than those on the Royal Society Corpus. It is mainly because that documents in a corpus with long time span have better discrimination on temporal semantic relations. For another thing, the overlapping degree of document representation clusters of BERT is more than that of TicTac on both corpora. For example, the margin of the cluster Northern Qi, Southern Liang, and Tang by BERT is not as much clear as that by TicTac, which produces a tighter grouping of categories and clearer margins. Such visualization results illustrate the efficiency of TicTac in temporal category discrimination.

4.8 Case Study

We compare TicTac with RoBERTa which achieves the best performance among all baseline systems. Specifically, we can observe the fine-grained performances of the two models by temporal category on the two datasets by means of evaluation metrics on each category. Table 5 in Appendix A displays the performances of the two models on each category.

It can be seen from the table that, in the Royal Society Corpus, TicTac outperforms RoBERTa across most temporal categories except for the period 1680–1800. It is worth noting that, TicTac achieves the highest F1 score of 76.47% during 1860–1880. On the other hand, in the Twenty-Four Histories Corpus, the performance gap between the two models is more distinct. For example, RoBERTa outperforms TicTac in the temporal category of Western Han, Southern Liang, and Northern Qi. However, TicTac outperforms RoBERTa in other periods. For example, in the period Eastern Han and Southern Song, TicTac achieves an F1 score increase by 4.62% and 6.81% in comparison with RoBERTa, respectively. To sum up, TicTac demonstrates a more stable performance on the Royal Society Corpus, while its effectiveness on the Twenty-Four Histories Corpus is more dependent on the specific temporal categories.

Figure 4 in Appendix A displays the confusion matrices of TicTac on the test sets of the two corpora. In the Twenty-Four Histories Corpus, TicTac performs well on some temporal categories such as Western Jin, Tang, and Later Jin, with the accuracy scores exceeding 85% in most cases. We also notice that, higher confusion rates occur at categories like Southern Song and Southern Liang

with Northern Qi, probably due to unclear semantic intervals from these short time-span and adjacent periods. In the Royal Society Corpus, TicTac shows superior performances on the 1660–1680 and 1860–1880 periods, probably because they are respectively the start and the end of the temporal class sequences that are easier to be identified than the other categories. On the other hand, TicTac has lower performances on the period 1680–1700 and 1720–1740, possibly due to the limited data scale on these temporal categories.

5 Conclusion

In this paper, we propose TicTac, a novel approach to the ATD task that integrates time-aware contrastive learning and time-aware ordinal classification to jointly model ordinally relative and absolute temporal relationships in long time-span texts. Experiments on two diachronic corpora show that TicTac outperforms state-of-the-art baselines, demonstrating its effectiveness in capturing temporal implications. TicTac also shows the limitations on short time-span texts, and further evaluation on multilingual datasets is needed to assess its generalization ability.

Limitations

This study acknowledges two primary limitations. First, due to the limited availability of text-dating datasets, experiments were conducted exclusively on Chinese and English datasets, and the model's performance remains untested on datasets in other languages. To fully assess the generalizability of the model, further validation on datasets from additional languages is necessary. Second, the proposed method encounters difficulties when applied to tasks involving texts with minimal or no temporal span. This challenge stems from the model's reliance on the evolutionary properties of language for its modeling, a characteristic that becomes less pronounced in datasets with shorter or absent temporal spans. Future work should aim to refine the model to more effectively capture lexical changes in such datasets.

Acknowledgments

This work is supported by Research Foundation of the Laboratory of Language and Artificial Intelligence at Guangdong University of Foreign Studies(No. LAI202305), Graduate Research Innovation Foundation of Guangdong University for Foreign Studies(No. 25GWCXXM-069), and China Ministry of Education Foundation(No. 21YJC740058).

References

Enrique Amigo, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3949, Online. Association for Computational Linguistics.

Anaëlle Baledent, Nicolas Hiebel, and Gaël Lejeune. 2020. Dating Ancient texts: an Approach for Noisy French Documents. In *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, Series Dating Ancient texts: an Approach for Noisy French Documents, pages 17–21.

Sidsel Boldsen and Fredrik Wahlberg. 2021. Survey and reproduction of computational approaches to dating of historical texts. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 145–156. Linköping University Electronic Press, Sweden. Place: Reykjavik, Iceland (Online).

Shuyang Cao and Lu Wang. 2022. Time-aware Prompting for Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP2022*, Series Time-aware Prompting for Text Generation, pages 7231–7246. Type: Conference Paper.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A Simple Log-based Loss Function for Ordinal Text Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nathanael Chambers. 2012. Labeling Documents with Timestamps: Learning from their Time Expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics(ACL)*, Series Labeling Documents with Timestamps: Learning from their Time Expressions, pages 98–106. Type: Conference Paper.

Alina Maria Ciobanu, Anca Dinu, Liviu Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2013. Temporal classification for historical Romanian texts. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 102–106. Association for Computational Linguistics. Place: Sofia, Bulgaria.

Angelo Dalli. 2006. Temporal classification of text and automatic document dating. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 29–32.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? Automatically determining publication dates. In *String Processing and Information Retrieval: 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011. Proceedings 18*, pages 221–236. Springer.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2022. Embed2Detect: Temporally Clustered Embedded Words for Event Detection in Social Media. *Machine Learning*, (2022)111:49–87. Type: Journal Article.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

Nattiya Kanhabua and Kjetil Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *International conference on theory and practice of digital libraries*, pages 358–370. Springer.

Siva Rajesh Kasa, Aniket Goel, Karan Gupta, Sumegh Roychowdhury, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2024. Exploring Ordinality in Text Classification: A Comparative Study of Explicit and Implicit Techniques. *arXiv preprint arXiv:2405.11775*.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The royal society corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931.

Meiwei Li, Yunhui Qin, and Wei Huangfu. 2023. RoBERTa: An Efficient Dating Method of Ancient Chinese Texts. In *Chinese Lexical Semantics*, pages 293–301. Springer Nature Switzerland. Place: Cham.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural Language Inference in Context - Investigating Contextual Reasoning over Long Texts. In *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence*, Series Natural Language Inference in Context - Investigating Contextual Reasoning over Long Texts, pages 13388–13396. Type: Conference Paper.

Riccardo Massidda. 2020. rmassidda@ DaDoEval: Document dating using sentence embeddings at EVALITA 2020. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 403.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 17–21.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

Octavian Popescu and Carlo Strapparava. 2015. SemEval 2015, Task 7: Diachronic Text Evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation(SemEval)*, Series SemEval 2015, Task 7: Diachronic Text Evaluation, pages 870–878. Type: Conference Paper.

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. Background Summarization of Event Timelines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Series Background Summarization of Event Timelines, pages 8111–8136. Type: Conference Paper.

Swayambhu Nath Ray, Shib Sankar Dasgupta, and Partha Talukdar. 2018. AD3: Attentive Deep Document Dater. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1871–1880, Brussels, Belgium. Association for Computational Linguistics.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-Aware Language Modeling for Historical Text Dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656. Association for Computational Linguistics. Place: Singapore.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL)*, Series Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs, pages 8017–8026. Type: Conference Paper.

Jungbin Son and Alice Oh. 2023. Time-Aware Representation Learning for Time-Sensitive Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Series Time-Aware Representation Learning for Time-Sensitive Question Answering. Type: Conference Paper.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. Improving Temporal Generalization of Pre-trained Language Models with Lexical Semantic Change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zuoyu Tian and Sandra Kübler. 2021. Period Classification in Chinese Historical Texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 168–177. Association for Computational Linguistics. Place: Punta Cana, Dominican Republic (online).

Jordi de La Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154. Publisher: Elsevier.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal Reasoning in Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Series Temporal Reasoning in Natural Language Inference. Type: Conference Paper.

Jiexin Wang, Adam Jatowt, and Yi Cai. 2024. Towards Effective Time-Aware Language Representation: Exploring Enhanced Temporal Understanding in Language Models. *CoRR*, abs/2406.01863.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. BiTimeBERT: Extending Pre-Trained Language Representations with Bi-Temporal Information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 812–821, New York, NY, USA. Association for Computing Machinery.

Yuting Wei, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li, and Bin Wu. 2025. A diachronic language model for long-time span classical Chinese. *Information Processing & Management*, 62(1):103925.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Xuejin Yu and Wei Huangfu. 2019a. A Machine Learning Model for the Dating of Ancient Chinese Texts. In *2019 International Conference on Asian Language Processing (IALP)*, pages 115–120.

Xuejin Yu and Wei Huangfu. 2019b. A Machine Learning Model for the Dating of Ancient Chinese Texts. In *2019 International Conference on Asian Language Processing (IALP)*, pages 115–120.

Sergey Zinin and Yang Xu. 2020. Corpus of Chinese dynastic histories: Gender analysis over two millennia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 785–793, Marseille, France. European Language Resources Association.

A Appendix

Dataset	Prompt Text
RSC	Given the following text, you need to determine the period class it belongs to. For example, a text belongs to the period class 1600–1680 means that it was written between 1600 and 1680. There are 11 period classes: 1660–1680, 1680–1700, 1700–1720, 1720–1740, 1740–1760, 1760–1780, 1780–1800, 1800–1820, 1820–1840, 1840–1860 and 1860–1880. Please return only one period class according to the text without any explanation. The text is as follows: `{text}`
24 Histories	请根据给定的文本内容，判断该文本所属的年代类别。例如，如果一个文本写于唐代 (618年-907年)，那么该文本的年代类别为唐代。年代类别包括：西汉、东汉、西晋、南朝宋、南朝梁、北朝齐、唐、后晋、宋、元、明、清。请只返回一个年代类别，不要添加任何额外的解释：`{text}`

Table 4: Prompt Templates for Twenty-Four Histories Corpus and Royal Society Corpus Datasets

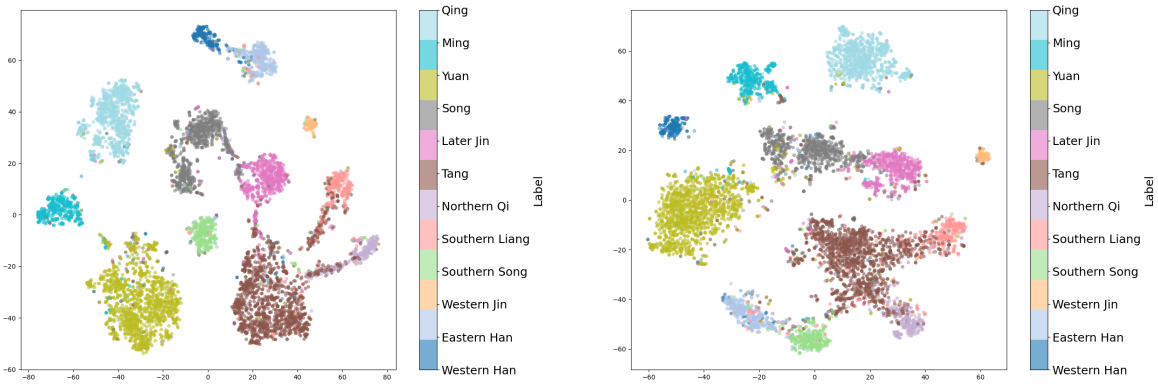

Figure 2: t-SNE plot of the CLS embeddings learned on the Twenty-Four Histories Corpus test set, consisting of 6911 samples and 12 categories. The plot on the left represents the TicTac model, while the plot on the right represents the BERT model.

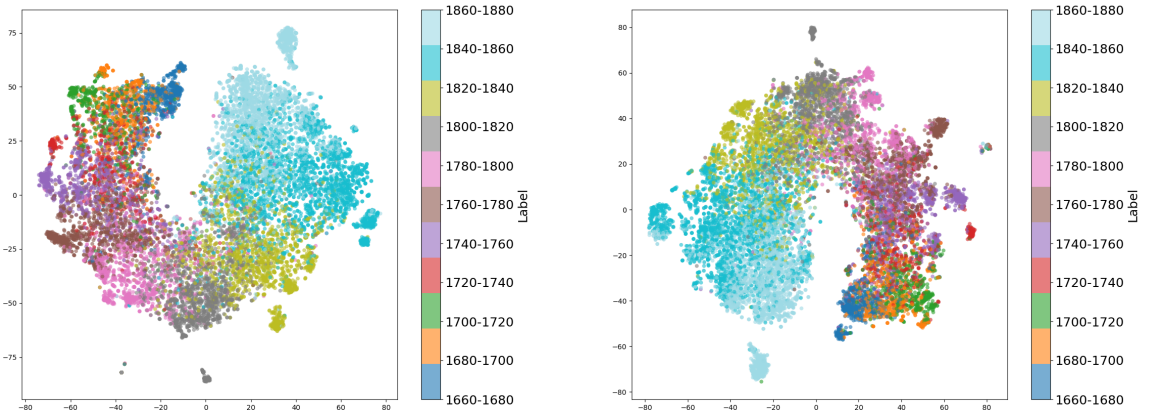

Figure 3: t-SNE plot of the CLS embeddings learned on the Royal Society Corpus test set, consisting of 13465 samples and 11 categories. The plot on the left represents the TicTac model, while the plot on the right represents the BERT model.

Dataset	Period	TicTac			RoBERTa		
		P	R	F1	P	R	F1
Twenty-Four Histories Corpus	Western Han	89.95	85.86	87.86	95.00	88.89	91.84
	Eastern Han	87.71	87.47	87.59	81.99	83.98	82.97
	Western Jin	90.80	95.18	92.94	93.18	91.79	92.48
	Southern Song	89.66	84.69	87.10	74.40	87.20	80.29
	Southern Liang	79.41	71.81	75.42	82.17	83.23	82.69
	Northern Qi	79.76	74.44	77.01	86.71	78.48	82.39
	Tang	86.04	90.95	88.43	88.08	87.89	87.99
	Later Jin	90.76	86.60	88.63	84.41	91.78	87.94
	Song	91.32	89.35	90.33	94.58	81.95	87.81
	Yuan	93.86	97.58	95.69	91.50	96.30	93.84
	Ming	93.40	91.24	92.31	87.26	94.48	90.73
	Qing	97.38	96.78	97.08	98.03	90.75	94.25
Royal Society Corpus	1660-1680	62.97	80.40	70.62	69.40	71.00	70.20
	1680-1700	48.48	36.62	41.72	47.53	45.81	46.66
	1700-1720	59.57	51.51	55.25	52.06	55.02	53.50
	1720-1740	53.53	31.39	39.57	38.10	40.60	39.31
	1740-1760	63.01	60.22	61.58	60.32	56.54	58.37
	1760-1780	57.89	74.73	65.24	61.68	61.45	61.57
	1780-1800	66.70	62.20	64.37	65.74	54.01	59.30
	1800-1820	76.45	65.94	70.81	62.54	78.42	69.59
	1820-1840	66.36	73.22	69.62	64.33	63.60	63.96
	1840-1860	68.37	79.11	73.35	66.94	63.98	65.42
	1860-1880	82.25	71.45	76.47	72.12	71.19	71.65

Table 5: Performance by historical period on TicTac and RoBERTa, evaluated by P, R, and F1

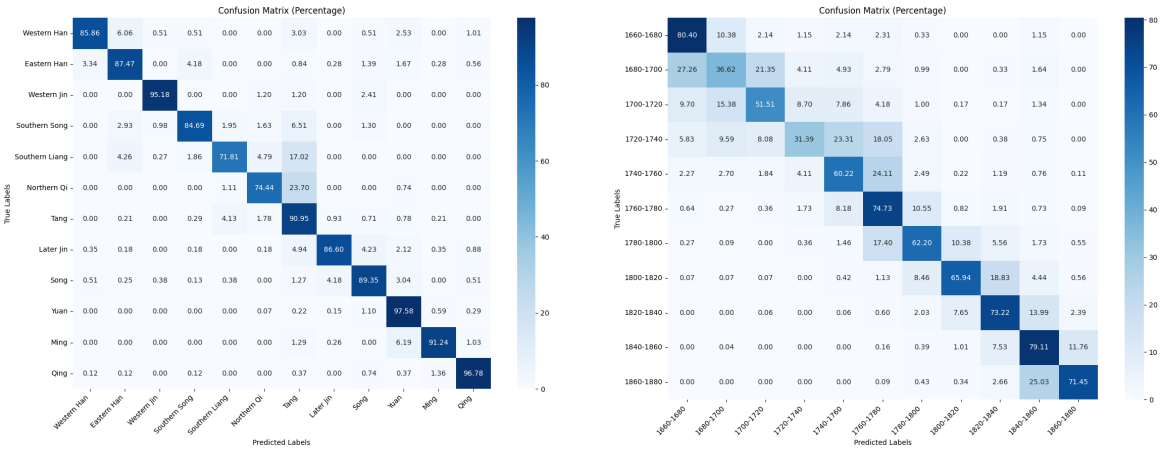

Figure 4: The confusion matrix of our model on the Twenty-Four Histories Corpus (left) and Royal Society Corpus (right). Rows represent true labels, and columns represent predicted labels. The percentage values in each cell indicate the proportion of instances correctly or incorrectly classified into each time period.