Research Article

A Time-Critical Topic Model for Predicting the Survival Time of Sepsis Patients

Wenping Guo,1,2 Zhuoming Xu,2 Xijian Ye,3 Shiqing Zhang,1 Xiaoming Zhao,1 and Xue Li4,5

1Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China
2College of Computer and Information, Hohai University, Nanjing 210098, China
3Polytechnic Institute, Zhejiang University, Hangzhou 310015, China
4Neusoft Institute of Information, Dalian Neusoft University of Information, Dalian 116081, China
5School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia

Correspondence should be addressed to Wenping Guo; guowp@tzc.edu.cn

Received 19 March 2020; Revised 18 August 2020; Accepted 29 August 2020; Published 15 September 2020

1.Introduction

Predicting the survival time of patients is an active research area for both clinicians and scientists [1–5]. It can significantly contribute to making decisions about clinical treatment, allocation of medical resources, and hospice care for patients [1]. Sepsis is a disease of life-threatening organ dysfunction caused by a dysregulated host response to infection [6]. Without timely treatment, sepsis can rapidly lead to tissue damage, organ failure, and death. Common signs and symptoms include fever, increased heart rate, increased breathing rate, and confusion. In US health systems, the cost for patients with sepsis accounted for more than $20 billion (5.2%) of total US hospital costs in 2011 [7]. In 2001–2010, one in twenty deaths in England was associated with sepsis based on information recorded on death certificates [8]. Although clinicians have made efforts to improve sepsis patient survival time, the mortality rate of sepsis is still very high [9, 10]. Thus, accurate prediction of survival time for sepsis patients could help clinicians conduct prevention, provide early warning and effective treatment, and reduce the mortality rate. Unfortunately, the pathogenesis of sepsis remains unclear. Predicting the survival time for specific diseases, such as sepsis, is still a challenging problem.

To help clinicians understand the overall body situations of a patient, ICUs have introduced many different mechanisms for describing the patient’s body situation and progress in the ICU. Different severity of illness scores has been used to predict sepsis or mortality risk scores in the ICU. The most widely used score-based methods include the acute physiology and chronic health evaluation (APACHE III) [11], the simplified acute physiology score (SAPS II) [12], the modified early warning score (MEWS) [13], the sepsis-related organ failure assessment (SOFA) [14], and quick SOFA (qSOFA) [6]. In addition, Zhang and Hong [15] proposed a novel score for predicting hospital mortality for severe sepsis. These score-based methods utilize a set of easily obtainable measurements from various patients to
generate risk scores. Although they allow clinicians to make rapid diagnostics of a patient, the obtained results are not satisfactory [16, 17]. Additionally, these score-based approaches only evaluate the patient body situation at specific times and cannot predict the survival time of sepsis patients because the development of sepsis is a time-sensitive process.

Sepsis is a life-threatening condition that arises when the body’s response to infection causes injury to its own tissues and organs [18, 19]. It is difficult to predict the development of sepsis based on a small number of measurements. Conventional topic models such as latent Dirichlet allocation (LDA) are unsupervised machine learning methods that can recognize latent topic information in massive document collections [20, 21]. Lehman et al. [22] proposed a novel approach for ICU patient risk stratification using a topic model. Ghassemi et al. [23] proposed a mortality model using a topic model to predict in-hospital mortality, 30-day postdischarge mortality, and 1-year postdischarge mortality. Vairamani [24] proposed an approach for mortality prediction in ICU patients based on LDA. Zhang et al. [25] proposed a novel survival topic model inspired by LDA for trauma patients. However, these topic models overlook the useful sequential information for disease development, thereby reducing the prediction accuracy for sepsis patient survival time. Unfortunately, the development and treatment of sepsis is a time-critical process that has a high correlation between the order of words in measurement and notes.

To address this issue, this paper proposes a time-critical topic model (TiCTM) inspired by the LDA model to predict the survival time of sepsis patients. The proposed TiCTM approach takes into account the time dependency structure between notes, measurement, and survival time of a sepsis patient. We consider the time-critical dynamic process of sepsis patients as an approximately linear variation under clinician treatment. Therefore, the linear change in the parameter of TiCTM to reflect this time-critical dynamic process, whereas the parameter of LDA is fixed.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed TiCTM for predicting survival time for adult sepsis patients in the ICU. The experiments and evaluations are discussed in Section 3. Finally, we conclude our research in Section 4.

2. Methodology

In this section, we first present a brief review of the classical LDA model and then describe our proposed TiCTM approach in detail.

2.1. Brief Review of LDA. LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [20]. The LDA model is represented as a probabilistic graphical model, as shown in Figure 1. The meaning of each notation for LDA is shown in Table 1.

The LDA model considers documents as the collection of words, which overlook the order of words. To address this issue, we propose a time-critical topic model inspired by LDA to predict the survival time of sepsis patients, as described as follows.

2.2. Our Proposed Method. Our proposed TiCTM approach considers the time dependency structure between notes, measurement, and survival time of a sepsis patient. The TiCTM model is depicted in Figure 2.

Assume that sequential notes and measurement sub-model have several phases. For phase 1, the meaning of each notation is shown in Table 2.

For phase $i \in \{2, 3, \ldots, I - 1\}$, the meaning of each notation is shown in Table 3.

For phase $I$, the meaning of each notation is shown in Table 4.

As shown in Figure 2, the proposed TiCTM model consists of two submodels: sequential notes and measurement and survival time prediction. This is because the development of sepsis is a time-critical process. The main idea of the TiCTM model is that we consider the time-critical dynamic process of sepsis patients as approximately linear variation under clinician treatment. Therefore, we employ the linear change in the parameter of TiCTM to reflect this time-critical dynamic process, whereas the parameter of LDA is fixed.

2.2.1. Sequential Notes and Measurement. For a sepsis patient $m$, the parameter $\mu_{im}$ represents the patient’s initial body situation. In everyday clinician treatment, the patient’s body situation changes $\Delta \mu$. Therefore, after $i_m$ days of clinician treatment, the parameter $\mu_{im}$ of patient $m$ will transit to $\mu_{im} + i_m \Delta \mu$, which is used to reflect the sequential changing process. The parameter $\mu_{im} + i_m \Delta \mu$ determines the disease probability distribution $\theta_i$. The parameter $\theta_i$ determines the disease $z_i$ for patient $m$. $\beta$ (probability of bigrams appearing in the topics of notes dictionary) and $z_i$ (topic variable) generate $N_{im}$ words in the notes of patient $m$. $\lambda$ (probability of measurements appearing in the topics of measurement dictionary) and $z_i$ generate $R_{im}$ measurements in the measurement result of patient $m$.

For each word $w_{i,n}$, $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, I$, we draw a topic assignment $z_{i,n} | \theta_i \sim \prod (A(\theta_i))$, where

$$A(\theta_i) = \left( \frac{1}{\sqrt{\pi}} e^{-\theta_i^2} \right)^{w_{i,n}},$$

where $A(\theta_i)$ represents the probability of topic appearing, which is determined by parameter $\theta_i$.

Draw a word $w_{i,n} | z_{i,n}, \beta \sim \prod (B(\beta))$, where

$$B(\beta) = \left( \frac{1}{\sqrt{\pi}} e^{-\beta^2} \right)^{w_{i,n}},$$

2 Scientific Programming
2.2.2 Survival Time Prediction. For each patient \( m \), the objective of the problem is to minimize the difference (Diff) between \( f_m \) and \( \bar{f}_m \):

\[
\min \text{Diff} = \min \sum_{m=1}^{M} \frac{1}{1 + e^{f_m}} \times (f_m - \bar{f}_m)^2.
\]

We hypothesize \( T_m \) as the real survival time for patient \( m \) after the second phase. \( f_m \) is the function for real survival time analysis \( T_m \).

Assume that patient \( m \) has the same probability \( p_m \) to survive every day; then, we can calculate the probability of death on the \((T_m + 1)\) day using

\[
p_m^{T_m} (1 - p_m), \quad p_m \in [0, 1].
\]

To maximize \( p_m^{T_m} (1 - p_m) \), it must be satisfied with

\[
p_m = \frac{T_m}{T_m + 1}.
\]

where \( p_m \) is a form of sigmoid function as

\[
p_m = \frac{1}{1 + e^{-x}}.
\]

To make \( p_m = 0 \) (Death) and \( p_m = 1 \) (Survive), we modify \( p_m \) to

\[
p_m = \frac{1.01}{1 + e^{-x}} - 0.005.
\]

Then, we can obtain formula (12) for \( f_m \) and formula (13) for \( T_m \):

\[
f_m = \log\left(\frac{1.01}{(T_m/(T_m + 1)) + 0.005} - 1\right).
\]

\[
T_m = \frac{p_m}{1 - p_m} = \frac{(1.01/(1 + e^{f_m}) - 0.005)}{1.005 - (1.01/(1 + e^{f_m}))}.
\]
2.3.2. Definition of the Survival Time Prediction Function.

To verify the effectiveness of the proposed TiCTM model, this paper uses the two-phase survival model to analyze survival time analysis for adult sepsis patients. The measurement and notes in the first phase limit the time for the patient’s admission within 24 hours. The second phase uses the patient’s measurement and notes after 24 hours to the last.

To calculate the probability of measurement \(x_1\) and note \(w_1\) in the first phase by body condition \(\mu_m\), we use

\[
\mu_m + \Delta \mu
\]

\[
\mu_m + (I - 1) \mu \Delta \mu
\]
Table 3: Notations for phase $i$.

| Notation | Description |
|----------|-------------|
| $i$      | The phase $i$ |
| $i_m$    | Patient $m$ treated by clinicians for $i_m$ days |
| $\Delta \mu$ | Parameter $\Delta \mu$ |
| $\theta_i$ | A $K$-dimensional normal random variable |
| $z_i$ | Topic variable |
| $w_{ij}$ | A set of $N_{ij}$ bigrams in a note |
| $N_{im}$ | Number of bigram results by $z_i$ |
| $x_i$ | Measurement $x_i$ |
| $R_{im}$ | Number of measurement results by $z_i$ |

Table 4: Notations for phase $M$.

| Notation | Description |
|----------|-------------|
| $M$      | Total number of patients |
| $l$      | Total number of phases |
| $I_m$    | Patient $m$ treated by clinicians for a total of $I_m$ days |
| $T_m$    | Real survival time for patient $m$ after phase $l$-1 |
| $\eta$ | Regression coefficient |
| $f_m$ | Function for computing real survival time |
| $f_M$ | Function for computing predicted survival time |
| Diff | The function value of the difference between $f_m$ and $f_M$ |

\[
p(\theta_1 | \mu_m) = \prod_{k=1}^{K} \frac{1}{\sqrt{\pi}} e^{-\left(\theta_{1k} - \mu_{mk} - i_m\Delta \mu\right)^2},
\]

\[
p(z_{2,n} | \theta_2) p(w_{2,n} | z_{2,n}; \beta) = \prod_{v=1}^{V} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{2,v}^2z_{2,v}^2 - \frac{1}{\sqrt{\pi}} e^{-\beta^2z_{2,v}^2}} \right)^{w_{2,n}} 
\]

\[
p(z_{2,r} | \theta_2) p(x_{2,r} | z_{2,r}, \lambda) = \prod_{v'=1}^{v'} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{2,v'}^2z_{2,v'}^2 - \frac{1}{\sqrt{\pi}} e^{-\lambda^2z_{2,v'}^2}} \right)^{x_{2,r}}.
\]

The likelihood function of the two-phase of patient $m$ can be obtained from

\[
L(w_1, w_2, x_1, x_2 | \mu_m, \Delta \mu, i_m, \beta, \lambda) = \int p(\theta_1 | \mu_m) 
\]

\[
\times \left( \prod_{n=1}^{N_m} \sum_{z_{1,n}=1}^{K} p(z_{1,n} | \theta_1) p(w_{1,n} | z_{1,n}; \beta) \right) d\theta_1,
\]

where

\[
p(z_{1,n} | \theta_1) p(w_{1,n} | z_{1,n}; \beta) = \prod_{v=1}^{V} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{1,v}^2z_{1,v}^2 - \frac{1}{\sqrt{\pi}} e^{-\beta^2z_{1,v}^2}} \right)^{w_{1,n}} 
\]

\[
p(z_{1,r} | \theta_1) p(x_{1,r} | z_{1,r}, \lambda) = \prod_{v'=1}^{v'} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{1,v'}^2z_{1,v'}^2 - \frac{1}{\sqrt{\pi}} e^{-\lambda^2z_{1,v'}^2}} \right)^{x_{1,r}}.
\]

To calculate the probability of measurement $x_2$ and note $w_2$ in the second phase by body condition $(\mu_m + i_m\Delta \mu)$, we use

\[
p(\theta_2 | \mu_m + i_m\Delta \mu) = \prod_{k=1}^{K} \frac{1}{\sqrt{\pi}} e^{-\left(\theta_{2k} - \mu_{mk} - i_m\Delta \mu\right)^2},
\]

\[
p(z_{2,n} | \theta_2) p(w_{2,n} | z_{2,n}; \beta) = \prod_{v=1}^{V} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{2,v}^2z_{2,v}^2 - \frac{1}{\sqrt{\pi}} e^{-\beta^2z_{2,v}^2}} \right)^{w_{2,n}} 
\]

\[
p(z_{2,r} | \theta_2) p(x_{2,r} | z_{2,r}, \lambda) = \prod_{v'=1}^{v'} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{2,v'}^2z_{2,v'}^2 - \frac{1}{\sqrt{\pi}} e^{-\lambda^2z_{2,v'}^2}} \right)^{x_{2,r}}.
\]

The log of the likelihood function of the two-phase of patient $m$ can be obtained from

\[
L(w_1, w_2, x_1, x_2 | \mu_m, \Delta \mu, i_m, \beta, \lambda) = \int p(\theta_1 | \mu_m) 
\]

\[
\times \left( \prod_{n=1}^{N_m} \sum_{z_{1,n}=1}^{K} p(z_{1,n} | \theta_1) p(w_{1,n} | z_{1,n}; \beta) \right) d\theta_1,
\]

where

\[
p(z_{1,n} | \theta_1) p(w_{1,n} | z_{1,n}; \beta) = \prod_{v=1}^{V} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{1,v}^2z_{1,v}^2 - \frac{1}{\sqrt{\pi}} e^{-\beta^2z_{1,v}^2}} \right)^{w_{1,n}} 
\]

\[
p(z_{1,r} | \theta_1) p(x_{1,r} | z_{1,r}, \lambda) = \prod_{v'=1}^{v'} \left( \frac{1}{\sqrt{\pi}} e^{-\theta_{1,v'}^2z_{1,v'}^2 - \frac{1}{\sqrt{\pi}} e^{-\lambda^2z_{1,v'}^2}} \right)^{x_{1,r}}.
\]
\[
\log(L(w_1, w_2, x_1, x_2 \mid \mu_m, \Delta \mu, i_m, \beta, \lambda)) \geq \left(-N_m - R_m - \frac{K}{2}\right) \log(\pi)
+ (N_m + R_m)\log K - \sum_{k=1}^{K} \mu_{m,k}^2 - \frac{1}{K} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \sum_{v=1}^{V} w_{1,n}^v \beta_{k,v}^2 - \frac{1}{K} \sum_{r=1}^{R_m} \sum_{k=1}^{K} \sum_{\nu=1}^{\nu'} x_{1,r}^\nu \lambda_{k,\nu}^2 + \log(c_m)
+ \left(-N'_m - R'_m - \frac{K}{2}\right) \log(\pi) + (N'_m + R'_m)\log K - \sum_{k=1}^{K} \left(\mu_{m,k} + i_m \Delta \mu_k\right)^2 - \frac{1}{K} \sum_{n=1}^{N'_m} \sum_{k=1}^{K} \sum_{v=1}^{V} w_{2,n}^v \beta_{k,v}^2 - \frac{1}{K} \sum_{r=1}^{R'_m} \sum_{k=1}^{K} \sum_{\nu=1}^{\nu'} x_{2,r}^\nu \lambda_{k,\nu}^2 + \log(c'_m).
\]

where

\[
\log(c_m) = \frac{K}{2} \log(\pi) - \frac{K}{2} \log\left(1 + \frac{N_m + R_m}{K}\right) + \frac{K}{K + N_m + R_m} \sum_{k=1}^{K} \mu_{m,k}^2
\]

\[
\log(c'_m) = \frac{K}{2} \log(\pi) - \frac{K}{2} \log\left(1 + \frac{N'_m + R'_m}{K}\right) + \frac{K}{K + N'_m + R'_m} \sum_{k=1}^{K} \left(\mu_{m,k} + i_m \Delta \mu_k\right)^2.
\]

To simplify the calculation, we use formula (21) to replace the log of the likelihood function of formula (19):

\[
\left(-N_m - R_m - \frac{K}{2}\right) \log(\pi) + (N_m + R_m)\log K - \sum_{k=1}^{K} \mu_{m,k}^2 - \frac{1}{K} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \sum_{v=1}^{V} w_{1,n}^v \beta_{k,v}^2 - \frac{1}{K} \sum_{r=1}^{R_m} \sum_{k=1}^{K} \sum_{\nu=1}^{\nu'} x_{1,r}^\nu \lambda_{k,\nu}^2 + \log(c_m)
+ \left(-N'_m - R'_m - \frac{K}{2}\right) \log(\pi) + (N'_m + R'_m)\log K - \sum_{k=1}^{K} \left(\mu_{m,k} + i_m \Delta \mu_k\right)^2 - \frac{1}{K} \sum_{n=1}^{N'_m} \sum_{k=1}^{K} \sum_{v=1}^{V} w_{2,n}^v \beta_{k,v}^2 - \frac{1}{K} \sum_{r=1}^{R'_m} \sum_{k=1}^{K} \sum_{\nu=1}^{\nu'} x_{2,r}^\nu \lambda_{k,\nu}^2 + \log(c'_m).
\]

After removing the constant term in formula (21), the log of the total likelihood function of the two-phase for all patients can be obtained from the following formula:

\[
\log\left(\prod_{m=1}^{M} L(w_1, w_2, x_1, x_2 \mid \mu_m, \Delta \mu, i_m, \beta, \lambda)\right) \sim - \sum_{m=1}^{M} \left(\frac{N_m + R_m}{K + N_m + R_m} \sum_{k=1}^{K} \mu_{m,k}^2 + \frac{N'_m + R'_m}{K + N'_m + R'_m} \sum_{k=1}^{K} \left(\mu_{m,k} + i_m \Delta \mu_k\right)^2\right).
\]

To maximize the log of the total likelihood function, formula (23) is obtained as follows:

\[
\mu_{m,k} = \frac{-(N'_m + R'_m) \times (K + N_m + R_m) \times i_m \times \Delta \mu_k}{(N_m + R_m) \times (K + N_m + R_m) + (N'_m + R'_m) \times (K + N'_m + R'_m)}.
\]
Under the given constraints, \( \sum_{k=1}^{K} \Delta \mu_k = 1 \), the following formulas are obtained:

\[
\Delta \mu_k = \frac{1}{\phi_k \times \sum_{k=1}^{K} (1/\phi_k)}.
\]

\[
\phi_k = \sum_{m=1}^{M} \frac{1}{\eta \times \sum_{k=1}^{K} \left( \frac{(N_m' + R_m') \times (k + N_m + R_m) + (N_m' + R_m') \times (k + N_m + R_m)}{k + N_m + R_m} \right)}.
\]

A body condition transition diagram is shown in Figure 3. Therefore, we can obtain the patient’s body condition at discharge.

We hypothesize \( \tilde{T}_m \) as the predicted survival time for patient \( m \) after the second phase. \( \tilde{f}_m \) is a function for predicting survival time analysis \( \tilde{T}_m \).

We predict the probability \( (\tilde{p}_m) \) for patient \( m \) to survive with

\[
\tilde{p}_m = \frac{1.01}{1 + e^{\sum_{v=1}^{V} \eta_{m,v,k} \cdot h_{m,v,k}}} - 0.005,
\]

where \( h_{m,v} \) is the indicator function. When the notes of patient \( m \) contain the \( v \)-th bigram word of the notes dictionary, the \( h_{m,v} \) value is 1; otherwise, the \( h_{m,v} \) value is 0. \( h_{m,v} \) is the indicator function. When the measurement of patient \( m \) contains the \( v \)-th measurement of the measurement dictionary, the \( h_{m,v} \) value is 1; otherwise, the \( h_{m,v} \) value is 0. \( \eta_{v,k} \) is the regression coefficient. The \( v \)-th bigram word of the dictionary represents a danger factor when \( \eta_{v,k} > 0 \), which indicates a shorter survival time. Otherwise, the \( v \)-th measurement of the measurement dictionary represents a protective factor when \( \eta_{v,k} < 0 \), which indicates a longer survival time. \( \eta_{v,k} \) is the regression coefficient. The \( v \)-th measurement of the measurement dictionary represents a danger factor when \( \eta_{v,k} > 0 \), which indicates a shorter survival time. Otherwise, the \( v \)-th measurement of the measurement dictionary represents a protective factor when \( \eta_{v,k} < 0 \), which indicates a longer survival time.

Then, we can obtain formula (27) for \( \tilde{f}_m \) and formula (28) for \( \tilde{T}_m \):

\[ \tilde{f}_m = \frac{1}{1 + e^{\sum_{v=1}^{V} \eta_{m,v,k} \cdot h_{m,v,k}}} \] (27)

\[ \tilde{T}_m = \frac{\tilde{p}_m}{1 - \tilde{p}_m} = \frac{(1.01 \left( 1 + e^{\tilde{f}_m} \right) - 0.005}{1.005 - \left( 1.01 \left( 1 + e^{\tilde{f}_m} \right) \right)} \] (28)

2.3.3. Object Optimization. To minimize the difference between \( f_m \) and \( \tilde{f}_m \), we define Diff as follows:

\[
\text{Diff} = \sum_{m=1}^{M} \frac{1}{1 + e^{\tilde{f}_m}} \times \left( f_m - \tilde{f}_m \right)^2 = \sum_{m=1}^{M} \frac{1}{1 + e^{\tilde{f}_m}} \times \left( \sum_{k=1}^{K} \frac{1}{1 + e^{-\mu_{m,k} \cdot (\tilde{T}_m + \tilde{f}_m)}} \right)^2.
\]

(29)
To find the optimal $\eta$ to minimize Diff, we obtain

$$
\arg\min_{\eta} \sum_{m=1}^{M} \frac{1}{1 + e^{-f_m}} \times \left( f_m - \frac{1}{1 + e^{-\mu_{m,k} - (i_m + j_m) \Delta \mu_k}} \sum_{v=1}^{V} h_{m,v} \eta_{v,k} + \sum_{v' = 1}^{V} h_{m,v'} \eta_{v,k}' \right) \frac{2}{\sum_{m=1}^{M} h_{m,v} + \sum_{v' = 1}^{V} h_{m,v'}}. 
$$

(30)

Then, we obtain the gradient update as the following formulas:

$$
\frac{\partial \text{Diff}}{\eta_{v,k}} = \sum_{m=1}^{M} \frac{-2}{1 + e^{-f_m}} \times \left( f_m - \frac{1}{1 + e^{-\mu_{m,k} - (i_m + j_m) \Delta \mu_k}} \sum_{v=1}^{V} h_{m,v} \eta_{v,k} + \sum_{v' = 1}^{V} h_{m,v'} \eta_{v,k}' \right) \frac{1}{\sum_{v=1}^{V} h_{m,v} + \sum_{v' = 1}^{V} h_{m,v'}}. 
$$

(31)

$$
\frac{\partial \text{Diff}}{\eta_{v,k}'} = \sum_{m=1}^{M} \frac{-2}{1 + e^{-f_m}} \times \left( f_m - \frac{1}{1 + e^{-\mu_{m,k} - (i_m + j_m) \Delta \mu_k}} \sum_{v=1}^{V} h_{m,v} \eta_{v,k} + \sum_{v' = 1}^{V} h_{m,v'} \eta_{v,k}' \right) \frac{1}{\sum_{v=1}^{V} h_{m,v} + \sum_{v' = 1}^{V} h_{m,v'}}. 
$$

(32)

3. Experiments

3.1. Experimental Design

3.1.1. Evaluation Tasks

Task 1: Evaluating the performance of the TiCTM model with different topics.

Task 2: Performance comparisons of TiCTM with LDA and linear regression models.

Task 3: F1 comparisons for different survival days with TiCTM ($K = 5$).

Task 4: Comprehensive performance comparisons of TiCTM (different topics) with LDA ($K = 5, 10$) and linear regression models using ROC.
3.1.2. Baselines. In this experiment, we use two methods as our baselines: LDA [14] and linear regression. For the linear regression model, the meaning of each notation is shown in Table 5.

\[ T_m = \frac{1}{1 + e^{-f_m}} \]

where \( f_m \) is a function for predicting survival time analysis.

\[ T_m \cdot \hat{T}_m \] is defined as

\[ T_m \cdot \hat{T}_m = \frac{\sum_{k=1}^{K} \left( \sum_{v=1}^{V} a_{r,k} h_{m,v} + \sum_{v'=1}^{V'} a_{v',k} h_{m,v'} \right)}{b} \]

where \( a \) and \( b \) are regression coefficients.

The objective of survival time prediction for sepsis patients is to minimize the loss function between \( f_m \) and \( \hat{T}_m \):

\[ \min \text{Loss} = \min \sum_{m=1}^{M} (f_m - \hat{f}_m)^2. \]

Then, we obtain the gradient update as formulas (36), (37), and (38):

\[ \frac{\partial \text{Loss}}{\partial a_{r,k}} = -2 \sum_{m=1}^{M} (f_m - \hat{f}_m) h_{m,v} \]

\[ \frac{\partial \text{Loss}}{\partial a_{v',k}} = -2 \sum_{m=1}^{M} (f_m - \hat{f}_m) h_{m,v'} \]

\[ \frac{\partial \text{Loss}}{\partial b} = -2 \sum_{m=1}^{M} (f_m - \hat{f}_m) \]

3.1.3. Evaluation Criteria. To evaluate the performance of the proposed method, we use a 3-fold cross-validation scheme. Evaluation metrics, such as recall, F1, and FPR, are adopted in this paper. They are defined as follows:

Recall (TPR) = TP/(TP + FN).
Precison = TP/(TP + FP).
Accuracy = (TP + TN)/(TP + TN + FP + FN).
F1 = precision \times recall \times 2/(precision + recall).
FPR = FP/(FP + TN), where TP indicates true positive, which means predicting a survival time less than or equal to a given time, while the true survival time is less than or equal to a given time; FP indicates false positive, which means predicting a survival time less than or equal to a given time but the true survival time is greater than a given time; TN indicates true negative, which means predicting a survival time greater than a given time, while the true survival time is greater than a given time; and FN indicates false negative, which means predicting a survival time greater than a given time but the true survival time is less than or equal to a given time.

3.1.4. Dataset. The dataset used in the experiments is the public Medical Information Mart for Intensive Care (MIMIC-III) [26]. MIMIC-III has been widely used by 845 publications as of the end of August 2019 [27]. The version of MIMIC is MIMIC-III v1.4, which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. The data spanned from June 2001 to October 2012. We used a dataset of 2,487 deceased adult (age > 14) sepsis patient records from MIMIC-III. The data processing flowchart is shown in Figure 4.

Patient features include text data, class data, and numerical data.

(1) Text data processing: clinical notes are also text data. This paper considers the clinical notes as a set of bigrams. We calculate the TF-IDF value of every bigram after removing the stopwords. Then, we sort every bigram according to TF-IDF and select the top 3,000 bigrams as the words for the dictionary.

(2) Numerical data processing: to calculate the mean and standard deviation (std) for numerical data, we divide it into five intervals: \((-\infty, \text{mean} - 1.5 \cdot \text{std}), \text{[mean} - 1.5 \cdot \text{std}, \text{mean} - 0.5 \cdot \text{std}), \text{[mean} - 0.5 \cdot \text{std}, \text{mean} + 0.5 \cdot \text{std}), \text{[mean} + 0.5 \cdot \text{std}, \text{mean} + 1.5 \cdot \text{std}), \text{[mean} + 1.5 \cdot \text{std}, +\infty)\). Each interval is used as a word in the measurement dictionary.

(3) Class data processing: each class of data is used as a word in the measurement dictionary.

3.2. Experimental Results

3.2.1. Evaluating the Performance of the TiCTM Model with Different Topics. To evaluate the performance of our proposed TiCTM model, we choose different topics, that is, \( K = 5, 10, 15, 20 \), to predict the survival time for adult sepsis patients in the ICU. The results are shown in Tables 6–9. In Table 9, we can see that 5 topics (\( K = 5 \)) achieve the best performance for predicting the probability for 30-day survival time. The number of topics is chosen by a 3-fold cross-validation scheme. We use the average value of 3-fold to select the best number of topics. It can be seen that the best number of topics for sepsis patients is 5 topics. The F1 score of 5 topics improved by 3.55% (in-hospital death), 3.88% (7 days), 2.56% (14 days), and 1.99% (30 days) compared to 20 topics. The possible reason is that there are many words in the dictionary, and the combination of topics and bigrams increases sharply when the number of topics increases. This situation will lead to model overfitting, and the F1 score will decrease.
3.2.2. Performance Comparisons of TiCTM with LDA and Linear Regression Model. Performance comparisons of TiCTM ($K=5, 10$) with LDA ($K=5, 10$) and the linear regression model are shown in Tables 10–13. The performance of our proposed TiCTM outperforms LDA and the linear model overall in terms of recall, precision, accuracy, and $F_1$-measure. When we consider sequential information for predicting the survival time of sepsis patients, the prediction accuracy increases.
3.2.3. F1 Comparisons for Different Survival Days with TiCTM \((K = 5)\). When predicting in-hospital death, 7-day survival, 14-day survival, and 30-day survival, the F1-score will increase, as shown in Table 14. For longer survival times, the patient’s condition remains steady. Therefore, the clinician intervenes less. Then, the clinician intervenes less. Therefore, the F1 score is more accurate for predicting longer survival times.

3.2.4. Comprehensive Performance Comparisons Using ROC. Figure 5 presents comprehensive performance comparisons of TiCTM (different topics) with LDA \((K = 5, 10)\) and the linear regression model using ROC. In Figure 5, we can see that the performance of our proposed TiCTM is better than that of the classic LDA and linear regression models.

### Table 6: In-hospital death.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.6492 | 0.5773    | 0.7186   | 0.6097|
| TiCTM \((K = 10)\) | 0.6190 | 0.5747    | 0.7138   | 0.5953|
| TiCTM \((K = 15)\) | 0.5901 | 0.5615    | 0.7027   | 0.5746|
| TiCTM \((K = 20)\) | 0.5895 | 0.5611    | 0.7019   | 0.5742|

### Table 7: Survival 7 days after discharge.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.8174 | 0.5360    | 0.6725   | 0.6447|
| TiCTM \((K = 10)\) | 0.7743 | 0.5250    | 0.6582   | 0.6225|
| TiCTM \((K = 15)\) | 0.7537 | 0.5169    | 0.6502   | 0.6111|
| TiCTM \((K = 20)\) | 0.7421 | 0.5147    | 0.6478   | 0.6059|

### Table 8: Survival 14 days after discharge.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.8489 | 0.5263    | 0.6443   | 0.6472|
| TiCTM \((K = 10)\) | 0.8179 | 0.5223    | 0.6379   | 0.6351|
| TiCTM \((K = 15)\) | 0.8032 | 0.5209    | 0.6355   | 0.6294|
| TiCTM \((K = 20)\) | 0.7857 | 0.5168    | 0.6311   | 0.6216|

### Table 9: Survival 30 days after discharge.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.8786 | 0.5395    | 0.6276   | 0.6655|
| TiCTM \((K = 10)\) | 0.8562 | 0.5419    | 0.6288   | 0.6608|
| TiCTM \((K = 15)\) | 0.8414 | 0.5372    | 0.6228   | 0.6532|
| TiCTM \((K = 20)\) | 0.8275 | 0.5328    | 0.6164   | 0.6456|

### Table 10: In-hospital death.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.6492 | 0.5773    | 0.7186   | 0.6097|
| TiCTM \((K = 10)\) | 0.6190 | 0.5747    | 0.7138   | 0.5953|
| LDA \((K = 5)\) | 0.3421 | 0.6733    | 0.7174   | 0.4528|
| LDA \((K = 10)\) | 0.4264 | 0.6414    | 0.7202   | 0.5128|
| Linear regression | 0.4771 | 0.4723    | 0.6407   | 0.4736|

### Table 11: Survival 7 days after discharge.

| Methods        | Recall | Precision | Accuracy | F1    |
|----------------|--------|-----------|----------|-------|
| TiCTM \((K = 5)\) | 0.8174 | 0.5360    | 0.6725   | 0.6447|
| TiCTM \((K = 10)\) | 0.7743 | 0.5250    | 0.6582   | 0.6225|
| LDA \((K = 5)\) | 0.5295 | 0.6258    | 0.7098   | 0.5728|
| LDA \((K = 10)\) | 0.5662 | 0.5641    | 0.6800   | 0.5645|
| Linear regression | 0.5768 | 0.472     | 0.6085   | 0.5174|
Conclusions

In this paper, we propose a time-critical topic model (TiCTM), which combines a patient’s measurement with notes, to predict the survival time for adult sepsis patients in the ICU. We consider useful sequential information for predicting the survival time of sepsis patients, thereby increasing the prediction accuracy. Our experimental results show that the proposed TiCTM has the best performance when predicting the probability for 30-day survival time using 5 topics. In addition, the performance of our proposed TiCTM is better than that of the classic LDA and linear regression models. In the future, our study will focus on the explainable machine learning model for predicting survival time in the ICU [28–31].

Data Availability

The data used to support the findings of this study are available from https://mimic.physionet.org.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are grateful to the Institute of Big Data and Artificial Intelligence in Medicine, Taizhou University, for providing computing resources. This research was funded by the Key Projects of Zhejiang Province’s Educational Science Planning (2017SB068), the Science and Technology Program of Taizhou (1803gy07), the Humanities and Social Science Project of the Chinese Ministry of Education (20YJAZH033), and the National Natural Science Foundation of China (61976149).

References

[1] L. Zhou, J. Cui, J. Lu, B. Wee, and J. Zhao, “Prediction of survival time in advanced cancer: a prognostic scale for Chinese patients,” Journal of Pain and Symptom Management, vol. 38, pp. 578–586, 2009.

[2] M. Zhou, L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, “A texture feature ranking model for predicting
survival time of brain tumor patients,” in Proceedings of 2013 IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society, Manchester, United Kingdom, October 2013.

[3] M. Zhou, L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, “Neuromaging based survival time prediction of gbm patients using cnns from small data,” in Proceedings of 2019 IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society, Bari, Italy, October 2019.

[4] A. N. V. Dehkordi, A. Kamali-Asl, N. Wen, T. Mikkelson, I. J. Chetty, and H. Bagher-Ebadian, “DCE-MRI prediction of survival time for patients with glioblastoma multiforme: using an adaptive neuro-fuzzy-based model and nested model selection technique,” NMR in Biomedicine, vol. 30, Article ID e3739, 2017.

[5] D. McPherson, C. Griffiths, M. Williams et al., “Sepsis-associated mortality in England: an analysis of multiple cause of death data from 2001 to 2010,” BMJ Open, vol. 6, pp. 1–7, 2013.

[6] L. Ou, J. Chen, K. Hillman et al., “The impact of post-operative sepsis on mortality after hospital discharge among elective surgical patients: a population-based cohort study,” Critical Care, vol. 21, p. 34, 2017.

[7] J. C. Yebenes, J. C. Ruiz-Rodriguez, R. Ferrer et al., “SOCMIC (Catalonian critical care society) sepsis working group: epidemiology of sepsis in catalonia: analysis of incidence and outcomes in a European setting,” Annals of Intensive Care, vol. 7, p. 19, 2017.

[8] A. Kniaus, D. P. Wagner, E. A. Draper et al., “The Apache III prognostic system,” Chest, vol. 100, no. 6, pp. 1619–1636, 1991.

[9] Z. Shi, W. Zuo, S. Liang, X. Yue, and X. Li, “IDDSAM: an integrated disease diagnosis and severity assessment model for intensive care units,” IEEE Access, vol. 8, pp. 15423–15435, 2020.

[10] A. E. Johnson, T. J. Pollard, L. Shen et al., “MIMIC-III, a freely accessible critical care database,” Scientific Data, vol. 3, Article ID 160035, 2016.

[11] J.-L. Vincent, R. Moreno, J. Takala et al., “The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure,” Intensive Care Medicine, vol. 22, no. 7, pp. 707–710, 1996.

[12] J.-L. Vincent, R. Moreno, “Clinical review: scoring systems in the critically ill,” Critical Care, vol. 14, no. 2, p. 207, 2010.

[13] A. B. Nielsen, H. C. Thorsen-Meyer, K. Belling et al., “Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the danish national patient registry and electronic patient records,” The Lancet Digital Health, vol. 1, no. 2, pp. e78–e89, 2019.

[14] J.-L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, “Sepsis definitions: time for change,” The Lancet, vol. 381, no. 9866, pp. 774–775, 2013.

[15] L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, “Neuromaging based survival time prediction of gbm patients using cnns from small data,” in Proceedings of 2019 IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society, Bari, Italy, October 2019.

[16] A. N. V. Dehkordi, A. Kamali-Asl, N. Wen, T. Mikkelson, I. J. Chetty, and H. Bagher-Ebadian, “DCE-MRI prediction of survival time for patients with glioblastoma multiforme: using an adaptive neuro-fuzzy-based model and nested model selection technique,” NMR in Biomedicine, vol. 30, Article ID e3739, 2017.

[17] D. McPherson, C. Griffiths, M. Williams et al., “Sepsis-associated mortality in England: an analysis of multiple cause of death data from 2001 to 2010,” BMJ Open, vol. 6, pp. 1–7, 2013.

[18] L. Ou, J. Chen, K. Hillman et al., “The impact of post-operative sepsis on mortality after hospital discharge among elective surgical patients: a population-based cohort study,” Critical Care, vol. 21, p. 34, 2017.

[19] J. C. Yebenes, J. C. Ruiz-Rodriguez, R. Ferrer et al., “SOCMIC (Catalonian critical care society) sepsis working group: epidemiology of sepsis in catalonia: analysis of incidence and outcomes in a European setting,” Annals of Intensive Care, vol. 7, p. 19, 2017.

[20] A. Kniaus, D. P. Wagner, E. A. Draper et al., “The Apache III prognostic system,” Chest, vol. 100, no. 6, pp. 1619–1636, 1991.

[21] Z. Shi, W. Zuo, S. Liang, X. Yue, and X. Li, “IDDSAM: an integrated disease diagnosis and severity assessment model for intensive care units,” IEEE Access, vol. 8, pp. 15423–15435, 2020.