Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 1852447, 15 pages
https://doi.org/10.1155/2022/1852447

Research Article

Classification of FinTech Patents by Machine Learning and Deep Learning Reveals Trends of FinTech Development in China

Hao Wang,[1] Xizhuo Chen,[2] Jiangze Du ,[3] and Kin Keung Lai[4,5]

[1]*School of Public Finance and Public Administration, Jiangxi University of Finance and Economics, No. 169, East Shuanggang Road, Nanchang 330013, China*
[2]*School of Finance, Jiangxi University of Finance and Economics, No. 169, East Shuanggang Road, Nanchang 330013, China*
[3]*School of Finance, Jiangxi University of Finance and Economics, Research Centre of Financial Management and Risk Prevention, No. 169, East Shuanggang Road, Nanchang 330013, China*
[4]*International Business School, Shaanxi Normal University, No. 620 West Chang'an Street, Xi'an, China*
[5]*Department of Management Sciences, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong SAR, China*

Correspondence should be addressed to Jiangze Du; jiangze.du@hotmail.com

Received 11 February 2022; Accepted 8 June 2022; Published 8 July 2022

Academic Editor: Long Wang

The development of the financial industry and its integration with information technology have promoted FinTech innovation. China is a major contributor to FinTech innovation, but few studies have systematically summarized FinTech innovation and development in China from the perspective of patents. This lacuna is attributable to the lack of a generally accepted classification of FinTech patents and the unavailability of classified Chinese FinTech patent text data. To fill this research gap, we developed a classification of Chinese FinTech patents and manually annotated a set of patent texts to train machine learning and deep learning models to classify massive Chinese patent application data and identify different types of FinTech innovations. Among the evaluated models, the character-level convolutional neural network (CNN) model and BERT model classified FinTech innovation most accurately. We used the character-level CNN to classify 20,529 Chinese FinTech patent applications from 2013 to 2020. The classified dataset was used to briefly analyze the history of FinTech innovation development in China and its future prospects.

1. Introduction

The essence of FinTech is to apply new technology to the financial field [1]. Cutting-edge technology can fundamentally change financial services by reducing transaction costs and improving transaction convenience and security [2]. The rapid development of FinTech is altering the mode of operation of the traditional financial system, and the integration of new technologies and financial systems has boosted global economic development. The continuous expansion of the scale of the FinTech market is the result of investments in research and development. Payment institutions are continuing to expand the application scope of FinTech by using new technologies such as big data, Internet of Things, and artificial intelligence, promoting financial industry reform and improving the operating efficiency of the financial system while controlling for risks [3]. Most notably, FinTech enables new types of lenders outside the traditional banking system to enter the financial market [4] and allows financial intermediary services to incorporate users' digital footprints into credit calculations and default rate prediction [5].

The launch of Alipay in 2004 is regarded as the beginning of this new financial industry in China. Alipay and WeChat Pay have since become important platforms for the huge flow of mobile payment funds in China. In the first half of 2020, mobile payments in China reached 196.98 trillion yuan, up 18.61% year on year, ranking first in the world. Despite the widespread use of FinTech in China's financial industry, the relevant literature has mainly focused on explaining the meaning of FinTech, the reasons for its emergence, and the impact of its level of development on regional economic development and corporate financing

constraints [6, 7]. A systematic and accurate overview of research on FinTech innovation in China is therefore lacking. The present study fills this research gap by identifying and classifying Chinese FinTech patents to analyze the technology layout and development status of FinTech in China. After reviewing previous research on the distribution of FinTech patents in the US and Europe [8–10], we screened patent application data from 2013 (the first year of Internet finance in China [11]) to 2020 using machine learning technology to classify the patent text. Patents related to FinTech were identified and classified according to different financial industry applications. A variety of machine learning algorithms and deep learning algorithms for identifying and classifying Chinese FinTech patents were tested, and the model with the best classification performance for patent classification combined with text filtering and manual identification was selected.

The rest of this study is organized as follows. In Section 2, we report related research progress. In Section 3, we introduce the databases used. The experimental procedure is explained briefly in Section 4, and the traditional machine learning models and deep learning models used in this study are described in Sections 5 and 6. We enumerate the classification results and evaluate model metrics in Section 7. In Section 8, we analyze the development process and orientation of FinTech innovation in China from the perspective of patents. Finally, we summarize the results of this study and outline future research directions in Section 9.

2. Related Work

With the boom of FinTech development, FinTech-related patent applications and licenses continue to grow and are becoming an important channel for enterprises to enter new markets. New technologies based on big data, cloud computing, and artificial intelligence are maturing, finding wider application in the financial industry and attracting attention from financial and academic circles. Journals such as *Review of Financial Studies* and *Journal of Management Information Systems* have featured FinTech innovation, and research perspectives include technologies applied in FinTech [12–14], opportunities and challenges faced by the traditional financial industry [15, 16], and corresponding risk controls and market regulations [17–19].

In the literature, innovation development is typically systematically summarized using annotated patent datasets and text-based machine learning approaches to classify patent databases [20–25]. Text classification approaches are mainly divided into traditional machine learning and deep learning approaches. The first step in text classification is to select the text classification category. Then, a small, representative segment of the unclassified data is manually labeled and divided into a training set, verification set, and test set. The training set and validation set are used to train and validate the model, and the test set is used to verify the accuracy of the model. The text classification algorithm treats a string of text as a sequence $W = [w_1, w_2, \ldots, w_T]$ composed of words or characters. In the data to be classified, each piece of data belongs to a certain category a_i, and set $A = \{a_1, a_2, \ldots, a_K\}$ contains all classification categories a_i. The model is calculated by the a posteriori probability of each category in a given text W:

$$P(a_i|W) = P(a_i|w_1, w_2, w_3 \ldots, w_T), \quad i = 1, 2, 3 \ldots k. \tag{1}$$

From this, the category of text M is calculated:

$$a = \arg\max_{a_i} P(a_i|w_1, w_2, \ldots, w_T). \tag{2}$$

Finally, the model with the best classification performance is chosen for unclassified data classification. Figure 1 shows commonly used text classification models.

In recent years, more and more researchers have tried to study FinTech innovation from the perspective of patents. Zhao et al. [26] directly take the number of patents applied by financial institutions as the proxy variable of FinTech innovation. Cojoianu et al. [27] use patents filed by FinTech startups to measure their innovation. However, these methods can only measure the FinTech innovation carried out by the financial sectors and cannot cover other sectors. Lee and Sohn [28] found that FinTech-related technological innovation could be found in patents, and the theme of FinTech innovation could be mined based on text analysis. However, due to the lack of annotated patents and corresponding datasets, there are few studies on FinTech patent identification based on machine learning or deep learning models.

Among available analyses, Chen et al. [8] manually annotated 1800 FinTech patents (in seven categories such as data analysis) from a US patent dataset from the BDSS database and used English text-based machine learning approaches for classification. The machine learning methods comprised a series of algorithms, including a linear support vector machine (SVM). The best performance on the test set was attained by the linear SVM, Gaussian SVM, and neural network model ensemble classifier based on patent text (accuracy on the test set: 82.6%; $F1$ score: 76.3%). In the subsequent classification of unlabeled data, 6,511 FinTech-related patents were extracted, of which 2,588 were from individuals and 3,923 were from companies. Descriptive statistics revealed that the largest number of FinTech patent applications in the United States was related to network security (1,179).

Xu et al. [9] used a random forest algorithm to identify and classify US FinTech patents from the Lens database covering the years 2014–2018. Compared with Chen et al.'s classification of FinTech patents, Xu et al. added lending as one of the categories and removed pure technology categories such as P2P. After labeling samples in seven categories, the best classification algorithm achieved an accuracy of 71.67% on a test set of 1800 FinTech patents. The largest category of licensed FinTech patents was related to mobile payments (682).

Caragea et al. [10] classified a large sample of 3850 manually annotated FinTech patents (in five categories: insurance, fraud, data analysis, investment, and payment) from patent databases under the jurisdiction of US and European laws using BERT, CNN, LSTM, and other methods. Compared with the seven categories used by Chen et al., Caragea et al.'s classification was more in line

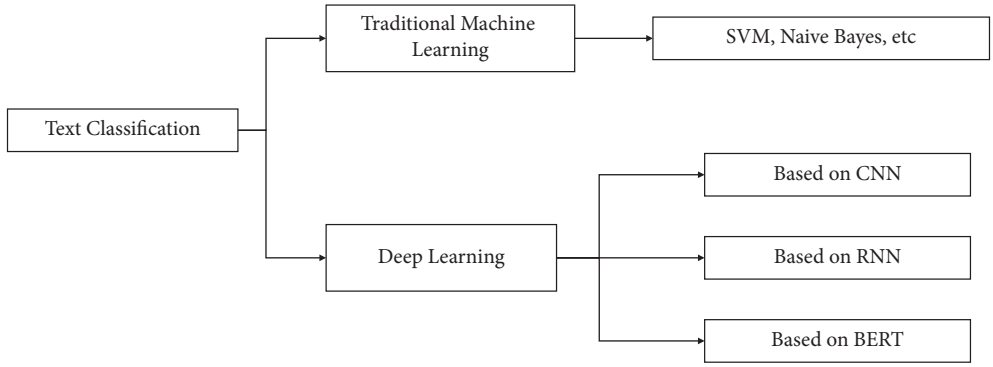

Figure 1: Commonly used text classification models.

with the definition of FinTech, as they abandoned the pure technology category and added insurance as a category to cover a large number of FinTech innovations in insurance and insurance risk prediction. Table 1 presents a comparison of the above three studies with our current work.

Chinese text classification has evolved from traditional machine learning algorithms to deep learning algorithms. Before classifying Chinese text, traditional machine learning and early deep learning algorithms first use classification modules such as Jieba to segment the text and remove stop words. Zhang et al. proposed character-level convolutional neural networks (ConvNets) for text classification. Comparisons of classification on a large number of datasets proved that the character-level CNN model performed better in the classification task than traditional word packets, n-grams, and TD-IDF variants. Based on this network structure, Google proposed the transformer structure and BERT model. The BERT model does not require word segmentation but operates directly on characters. For Chinese classification tasks, the word segmentation object in BERT is a single Chinese character.

3. Fintech Dataset

The patent application databases of various countries are not yet consolidated. Consequently, only the data for patents that are under the jurisdiction of the laws of a specific country and some patent data that are shared for the purpose of activity exchange can be retrieved from that country's official patent data. Previous studies of FinTech innovation have used patent data from the United States and Europe and omitted FinTech patents in China. Here, we used patent application data published by the China National Intellectual Property Administration from January 1, 2013, to December 31, 2020. Because FinTech development started relatively late in China compared with the United States, most patent applications have not yet been authorized, and patent application data are therefore used as the research object. Because an innovation subject applies for a patent once the innovation activities have achieved phased results, patent application data are more appropriate than patent licensing data for analyzing the boom in FinTech innovation in China.

The patent application dataset for China is very large, with more than 10 million patent applications in our selected time interval. To improve the classification efficiency, we referred to previous research on patent classification to preliminarily clean and filter the patent data. The official patent document classification and retrieval tool used by countries around the world is the International Patent Classification (IPC) compiled according to the Strasbourg Agreement on International Patent Classification signed in 1971. The major IPC categories are given in Table 2. The patent data of the China National Intellectual Property Administration also follow the IPC. In general, patents related to FinTech fall under categories *G* (Physics) and H (Electricity).

A preliminary screen of the 5,398,266 patent application data falling under categories G and H eliminated categories unrelated to finance or economics, such as nuclear physics. Then, we preliminarily used a machine learning algorithm to classify the patent data into finance-related and finance-independent patents after training the model on the general financial dataset. To ensure the integrity of the financial patents, we combined the patents that the different classification methods identified as potentially related to finance into a dataset containing 77,418 patent applications. This dataset is a collection of all patents roughly classified as related to FinTech. In the next steps, we used the manually annotated FinTech categories and different classification models to accurately classify the patent data.

4. Experimental Steps

FinTech encompasses the application of a variety of technologies in the financial industry. For example, information processing technology can accelerate the speed of information processing, enhance the ability of technology users to obtain information, reduce costs associated with the traditional reliance on human resources, and reduce redundant expenses. Although the technologies applied by different types of financial services overlap, their influences on China's financial development and corporate governance differ depending on the specific business context. In the earliest research on FinTech patents, Chen et al. divided FinTech patents into data analysis, Internet of Things, mobile payments, cybersecurity, blockchain, P2P, and robo-advising. Blockchain and Internet of Things do not fully

Table 1: Comparison of studies of the classification of FinTech patents.

	Dataset characteristics	Chen et al.	Xu et al.	Caragea et al.	Our work
1	Patent data source	BDSS	Lens	Orbis/Patsat	CNIPA
2	Year	2003–2017	2014–2018	2000–2017	2013–2020
3	Country	United States	America	America + Europe	China
4	Language	English	English	English	Chinese
5	Optimal classification algorithm	Ensemble classifier	Random forest	BERT	Character-level CNN
6	Number of patents manually labeled	1800	1800	3850	5400
7	Number of FinTech patents	6511	3602	25580	20529

Table 2: Major IPC categories.

Patent class	Meaning
A	Human necessities
B	Performing operations; transporting
C	Chemistry; metallurgy
D	Textiles; paper
E	Fixed constructions
F	Mechanical engineering; lighting; heating; weapons; blasting
G	Physics
H	Electricity

align with the general definition of FinTech, as these technologies are generally applied to specific financial industries such as insurance, lending, and data processing and thus have high overlap with other FinTech classification labels. Although there is no unified classification of FinTech patents, ongoing updates are gradually converging with the definition of FinTech. In this study, we referred to previous FinTech research [1–19] and research on the development of FinTech in China [29–31] that has analyzed and summarized the main characteristics of FinTech to divide the patent data into seven categories: FinTech-unrelated, payment, loan, insurance, security, data analysis, and investment. Table 3 provides the specific classifications.

Based on the categories in Table 3, representative patent texts were manually selected and labeled. The final labeled sample contained 5200 patents, including 2200 patents unrelated to FinTech and 3,000 FinTech-related patents (500 per category).

For a more intuitive presentation of the contents of FinTech patents, we randomly select one patent from each category as an example and report in Table 4. Meanwhile, considering that the patent abstract is too long to be a whole display, we extracted the contents that can reflect the characteristics and uses from the original abstract of patents and report in Introduction of Table 4.

The process of classification is illustrated schematically in Figure 2. The labeled sample was divided 8 : 1 : 1 into the training set, validation set, and test set, and traditional machine learning models (SVM, KNN, decision tree, and naive Bayes) and deep learning models (CNN, RNN, LSTM, and BERT) were used for text classification. To evaluate the performance of the different models in classifying Chinese patents, we used standard metrics such as precision (Pr), recall (Re), $F1$ score ($F1$), and accuracy (Acc):

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \\ \text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \end{aligned} \tag{3}$$

The above metrics take values between 0 and 1, where values closer to 1 indicate higher classification accuracy and a superior model. The parameters in the above formulas include true positive cases (TP), false positive cases (FP), false negative cases (FN), and true negative cases (TN). A TP occurs when the model correctly predicts a positive category sample, whereas a TN means that the model correctly predicts a negative category sample. An FP occurs when the model incorrectly predicts that a negative category sample is positive, while an FN corresponds to the incorrect prediction of a positive category sample as a negative category sample.

5. Traditional Machine Learning Models

This section describes the traditional machine learning models evaluated for patent classification. All models used are supervised machine learning models. The typical steps for text-processing are segmenting words, removing pause words and low-frequency words, performing feature selection, converting string text to text represented by vectors, and finally classifying the text. In English text, the spaces between words act as natural boundaries, and word segmentation is not necessary. However, Chinese text must be segmented. Automatic segmentation of a sentence into reasonable words can be performed by a computer following semantic logic. In natural language processing, the word is the smallest unit, and the accuracy of word segmentation directly affects the results of text classification.

Feature selection is required for Chinese word segmentation. If all feature words in the text are used to represent the text, the dimension of the feature space will typically be greater than 100,000. Such a high-dimensional space will greatly reduce the calculation efficiency or even make completion of the calculation impossible. Words must be selected from the text to form a new feature space and achieve dimensionality reduction; words with very weak

Table 3: FinTech categories in China.

FinTech category	Interpretation
Data analytics	The use of big data, cloud computing, and other pieces of information technology to clean, filter, and analyze large amounts of data related to the financial industry collected through various channels such as mobile banking
Lending	The use of information technology to deal with deposit and loan business
Insurance	The application of information processing technology for the intelligent recommendation of insurance products, policyholder risk identification, and other insurance businesses
Payment	The use of the Internet of Things and other technologies to process mobile payment business, provide payment function interfaces, and make payments intelligent, convenient, and cashless
Security	The application of fingerprint verification, iris verification, and face recognition technology in financial industry and equipment security management to improve the security of financial activities
Investment	Quantitative investment by enterprises and intelligent portfolio recommendation based on the analysis of investors' risk preferences

Table 4: Examples of FinTech patents.

Category	Patent	Applicant	Introduction
Data analytics	Abnormal warning methods, devices, systems, equipment, and media of big data products	Shenzhen Qianhai Micro Public Bank Co., Ltd.	The invention relates to the field of big data technology of FinTech and converts the error information of big data products into error codes and error descriptions for the convenience of operation and maintenance personnel to view and improve efficiency
Investment	Generating method and device of financial market product trading report	Bank of China Limited	The invention discloses a method and device for generating a financial market product trading report, in which the transaction information of all target customers is collected within a specified period
Security	Method and device of bank user authentication based on behavior characteristics	Industrial and Commercial Bank of China Limited	The invention proposes a bank user authentication method based on behavior characteristics, which makes full use of the user identity to identify the behavior characteristics of all kinds of business in the bank and labels the behavior characteristics
Insurance	Method, device, and storage medium for calculating insurance probability	Ping An Technology (Shenzhen) Co., Ltd.	The invention calculates the common characteristics of the customers who buy insurance through the GBDT model and then injects the customer data of the target customers into the model to obtain the insurance probability of the target customers
Lending	The invention relates to a loan system data processing method	China Construction Bank Corporation	The invention discloses a loan system data processing method, device, device, and storage medium. By dividing the prebatch and the primary batch, batch errors can be exposed in advance during prebatch, preventing dirty data from being directly stored in the database
Payment	Payment methods, devices, and systems	Tencent Technology (Shenzhen) Co., Ltd.	The invention discloses a payment method, equipment, and system. It solves the problem of complicated operation of online payment process when the third-party merchants access the shopping platform application in the form of web page

contributions to the text can be omitted. For example, the adverb "de" appears in almost all texts and is a frequent and meaningless feature that is usually deleted as a stop word.

Humans understand text in the form of character coding, whereas a computer system requires binary coding. Consequently, Chinese text must be transformed into binary code to allow the computer to calculate the text information. One commonly used text representation model is the vector space model. However, the weight of many feature words in vector space is zero, leading to a less-than-ideal classification effect. For traditional machine learning classification, we use TD-IDF text representation. TF-IDF is a statistical method that evaluates the importance of a single word in a dataset or a corpus. The importance of a word increases with the frequency of its appearance in a document but decreases with the frequency of its appearance in the corpus.

In this study, we evaluated the following traditional machine learning models: SVM, *k*-nearest neighbor (KNN), naive Bayes, and decision tree.

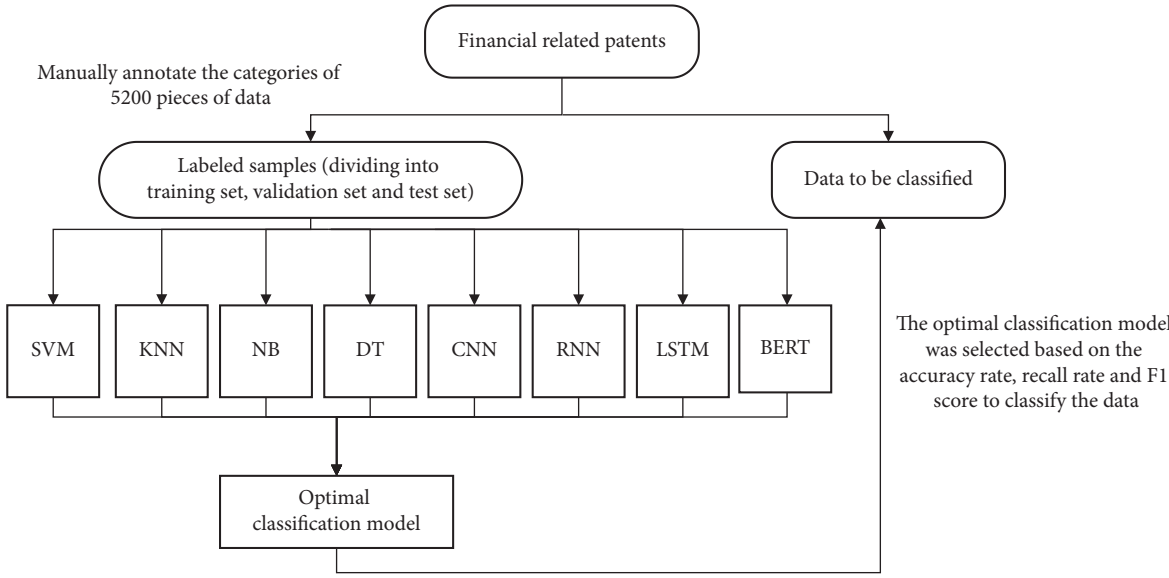

Figure 2: Patent data classification steps.

5.1. SVM. SVM is a kind of general feedforward neural network. The basic model finds the best separation hyperplane in the feature space that gives the largest interval between positive and negative samples in the sample training set. Depending on the sample's characteristics, the sample can be classified as linearly separable or linearly inseparable, as shown in Figure 3.

Linearly separable means that two kinds of samples can be separated by a straight line in a two-dimensional space. For a training sample $\{(x_i, d_i)\}$, where x_i is the ith sample of the input pattern and d_i is the corresponding expected response (target output), it is first assumed that the class represented by the subset $d_i = +1$ and the pattern represented by $d_i = -1$ are linearly separable. The hyperplane decision surface equation for separation is as follows:

$$f(x) = \omega^T x + b, \tag{4}$$

where x is the input vector, ω is the adjustable weight vector, and b is the offset. For a given weight vector ω and offset b, the goal of SVM is to find the hyperplane with the largest separation edge. Under this condition, the decision surface is called the optimal hyperplane.

A high-dimensional space can be separated by a high-dimensional function. Linearly inseparable means that the sample features are mapped to high-dimensional space by a Gaussian kernel function. Nonlinear features are transformed into linear separable features so that the sample can be processed by a linear separable method.

SVM can be used in a variety of supervised learning algorithms, including classification, regression, and anomaly detection, and has many advantages among traditional machine learning algorithms, such as high efficiency in high-dimensional space, a data dimension that is larger than the number of samples, and the ability to use a subset of the training set in the decision function (commonly known as a support vector), which can make efficient use of computer memory. However, SVM also has many disadvantages when it is used for a classification task. For example, if the number of features is much larger than the number of samples, overfitting can occur easily when selecting the kernel function for training. Overcoming this limitation usually requires regularization, among other means of dealing with overfitting.

5.2. KNN. KNN was first proposed by Cover and Hart in 1968. It is mature in both theory and research and is one of the simplest machine learning algorithms. The operation idea of this method is as follows: if most of the k-nearest samples in the feature space belong to a certain category, then the sample should also belong to this category. KNN determines the category of the samples to be classified based only on the category of the nearest one or several samples.

The drawback of this method is that the computational burden is very large, as the distance of each text to be classified from all known samples must be calculated to obtain the text's k-nearest neighbor points. The most common solution is to manually process the known sample points and remove the samples with little effect on classification in advance. The reverse KNN method can reduce the computational complexity of the KNN algorithm and improve the efficiency of classification. The KNN algorithm is more suitable for text classification of large sample sizes, as classification errors can occur easily when the sample size is small.

5.3. Naive Bayes. Bayesian methods use theories of probability and statistics to classify a sample dataset. Because the algorithm has a solid mathematical foundation, all Bayesian algorithms used for classification have higher accuracy and a lower error rate when the dataset is large. Bayesian methods combine prior probability and posterior probability to avoid the subjective bias of only using prior probability. Naive Bayes methods are a group of supervised learning algorithms

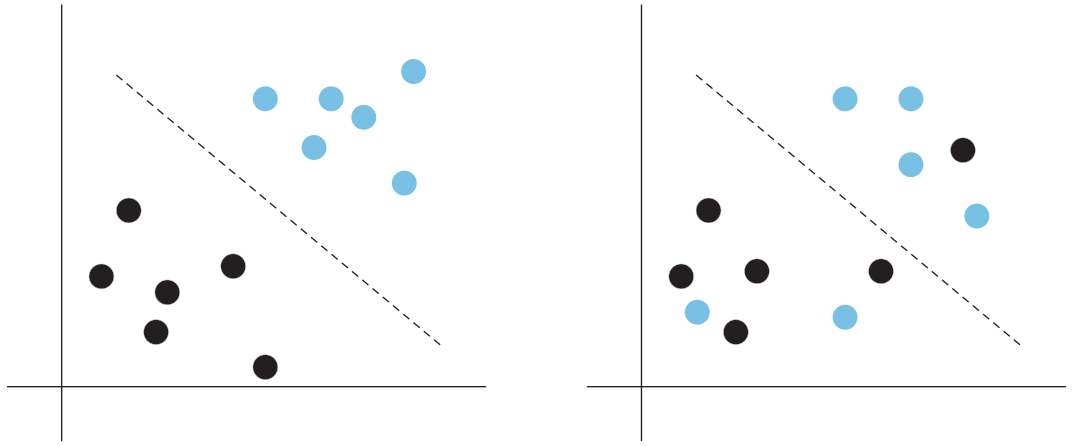

Figure 3: Linearly separable and linearly inseparable samples.

based on the Bayes theorem, which simply assumes that each pair of features is independent of each other. For a class y and x_1 to x_2 related eigenvectors of N, the naive assumption that each pair of features is independent of each other is used:

$$P(x_1|y, x_1, \ldots x_{i-1}, x_{i+1}, \ldots x_n) = P(x_i|y). \quad (5)$$

The maximum posterior probability can be used to estimate $P(x_i|y)$, which is the relative frequency of class y in the training set. The various naive Bayes classifiers differ mainly in the assumptions made when dealing with $P(x_i|y)$ distributions. Although the assumptions of the naive Bayes model are relatively simple, it works quite well in actual classification tasks, and only a random training sample is needed to estimate the required parameters. Naive Bayes learners and classifiers are much faster than other more complex methods. The decoupling of the conditional distribution of classification means that each feature can be estimated independently as a one-dimensional distribution. This in turn helps to mitigate the problems caused by dimensional disasters.

The Bayesian algorithm used in this study is a multinomial distributed naive Bayesian model, which is commonly used for text classification tasks and has shown excellent performance in previous experimental studies. The distribution parameter is determined by the $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ vector of each class y, where n is the number of features (for text classification, the number of features is the size of the vocabulary) and θ_{yi} is the probability $P(x_i|y)$ of feature i in the sample belonging to class y. θ_y is estimated using the smoothed maximum likelihood estimation method to calculate the relative frequency:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}. \quad (6)$$

In the above formula, $N_{yi} = \sum_{x \in T} x_i$ is the number of times that feature i in training set T appears in class y, and $N_y = \sum_{i=1}^{n} N_{yi}$ is the sum of all features appearing in class y. The prior smoothing factor $\alpha \geq 0$ is applied to features that do not appear in the learning sample to prevent the occurrence of a model calculation result with probability 0. $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

5.4. Decision Tree. In traditional machine learning, a decision tree is a prediction model that represents a mapping relationship between object attributes and object values. Each node in the decision tree represents a particular object, each bifurcation path represents a possible value, and each leaf is the value of an object represented by the path taken from the root node to that leaf. A decision tree has a single output. To output complex values, multiple independent decision trees can be built to output different values. The decision tree technique is frequently used in data mining and can be used for data classification, prediction, and regression. For the task of classification, the decision tree model is constructed according to the given dataset in the training stage, and the most valuable feature segmentation node is selected from the root node. The test phase and prediction phase are based on the decision tree model constructed by training.

6. Deep Learning Model

Four deep learning models are evaluated in this study for patent classification: CNN, RNN, LSTM, and BERT.

6.1. CNN. The convolutional neural network (CNN) is a representative deep learning algorithm and a type of feed-forward neural network that comprises convolutional computation and a deep structure. Because CNNs feature convolution kernel parameters within the neural network hidden layer and interlayer connection sharing of sparsity, learning pixel points and audio requires less computation, producing stable effects and data characteristics without additional requirements. Consequently, CNNs are widely used in image and text classification.

The CNN model used in this study is based on the implementation of TensorFlow on the Chinese dataset and classifies patent text based on character-level CNN. The

character-level CNN model encodes Chinese sentences into character sequences as input. Encoding is performed by specifying a Chinese word list of size M for input and quantifying each Chinese character using 1-of-M encoding ("one-hot" encoding). The sequence of characters is then converted into a sequence of m vectors of fixed length L. If the length of the paragraph is less than L, 0 is added after the sequence; if the paragraph length is longer than L, the field length greater than L is cutoff. Any character that is not in the word list (including whitespace characters) is encoded as a full zero vector. The character quantization order is reversed so that the most recent read of the character is always near the start of the output, which allows the full connection layer to associate weights with what the model reads.

The main operations of the model are shown in Figure 4. The first layer embeds text into a low-dimensional vector. The core of the model is the next layer, the CNN layer, which is the convolutional computing layer for calculating one-dimensional convolution. For example, assuming a discrete input function $g(x) \in [1, l] \longrightarrow \mathbb{R}$ and a discrete kernel function $f(x) \in [1, k] \longrightarrow \mathbb{R}$, the convolution $h(y)$ between f and g with stride d and migration constant c can be defined as

$$h(y) = \sum_{x=1}^{k} f(x) \cdot g(y \cdot d - x + c). \tag{7}$$

Next, we pool the maximum result calculated by the convolutional layer into a long feature vector and add dropout regularization. The convolution is again calculated given a discrete input function $g(x) \in [1, l] \longrightarrow \mathbb{R}$, stride d, and migration constant c. The max-pooling function $h(y)$ of $g(x)$ can be defined as

$$h(y) = \max_{x=1}^{k} g(y \cdot d - x + c). \tag{8}$$

The softmax layer is then used to classify the result.

6.2. RNN and LSTM. CNNs are unable to process sequence data due to the loss of sequence features during the pooling operation in the pooling layer. RNNs were introduced to alleviate this shortcoming. Text classification based on RNNs has developed rapidly in response to the need to process sequence data. As shown in Figure 5, the input vector of the neural network A containing several layers is x_t, and the output vector is h_t. The output calculated in the previous step is used as the input in the next step to form a chain structure of A through cyclic calculation.

However, RNNs inherit the flaw of deep architecture: the deeper the network, the more obvious the gradient explosion and gradient vanishing. Therefore, improvements of the structure of the RNN model have been proposed, such as LSTM (long short-term memory). As shown in Figure 6, in addition to the internal hidden state, LSTM adds a cell state. The information in the cell state is controlled by three hidden gates: the input gate, forget gate, and output gate. The forget gate determines what information to discard from the cell state, and the input gate determines how much new information to add to the cell state. This calculation directly affects the output cell state. The output gate determines the output value based on the cell state. These gates are used to determine the information in the previous cell that needs to be forgotten, update information, or output information for the current cell state.

6.3. BERT. BERT is a language model proposed based on Transformer structure that encodes tags and sentences into dense vector representations.

First, the BERT model needs to be pretrained on a large number of unlabeled language texts; the text marker sequence is used as the input. The first marker is a special marker represented by [CLS] (the output of (CLS) can be regarded as the semantic output of the whole text sequence), and the input sequence of BERT is composed of one or two sentences. Separated by another special tag [SEP], Transformer converts the embedding of the original input to the embedding of the context output. Figure 7 shows the operation structure of the BERT model when two sentences are entered. After preprocessing and fine-tuning the resulting model, the text-processing task is further pretrained to match the downstream task.

7. Classification Results

In this section, we present the classification performance of the different models and the classification results obtained using the optimal model. Table 5 provides the classification metrics for the different categories. The classification performances of character-level CNN and BERT were significantly better than those of the other classification models. As shown in Figure 8, the CNN had the best classification performance in terms of average score and Acc, and thus we chose CNN as the optimal model to analyze the data of 77,418 unclassified patents.

Applying the character-level CNN to patent text classification yielded a total of 20,529 FinTech patents. The steps of patent are given in Table 6.

After classifying 77,418 potentially FinTech-related patents and removing those belonging to the "FinTech-unrelated" category, we obtained a dataset containing 25,153 FinTech-related patents (belonging to six FinTech-related categories, such as payments). Removing FinTech patent applications filed abroad left a total of 20,529 FinTech patents. After excluding patents filed by individuals, universities, and so on, the total number of FinTech patents filed by companies (including banks) was 18,562. A total of 5,450 companies (including unlisted companies) applied for FinTech patents. A total of 81 banks (including branches) applied for 2009 FinTech patents.

8. Analysis of FinTech Patent Applications

In this section, we use the classified FinTech patents to briefly analyze the development of FinTech innovation in China. The dataset contains a total of 20,529 patents covering all FinTech patents filed in China between 2013 and 2020 based on classification by the deep learning model.

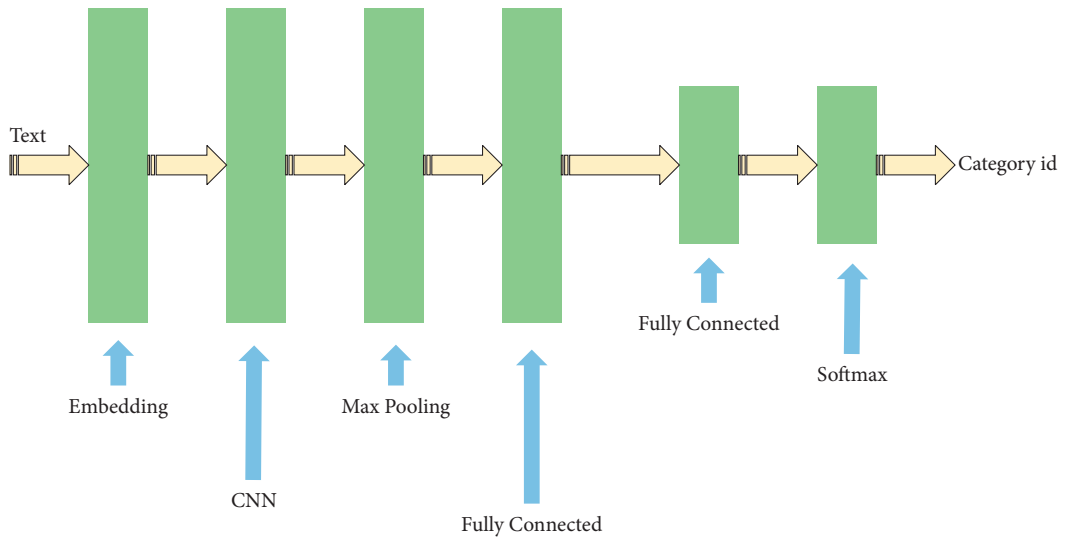

Figure 4: Convolutional neural networks for text classification.

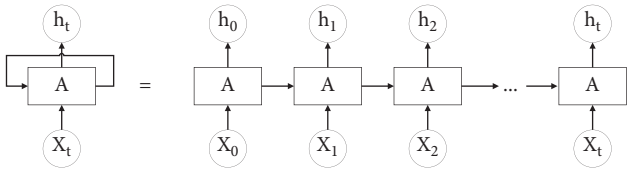

Figure 5: Recurrent neural network (RNN).

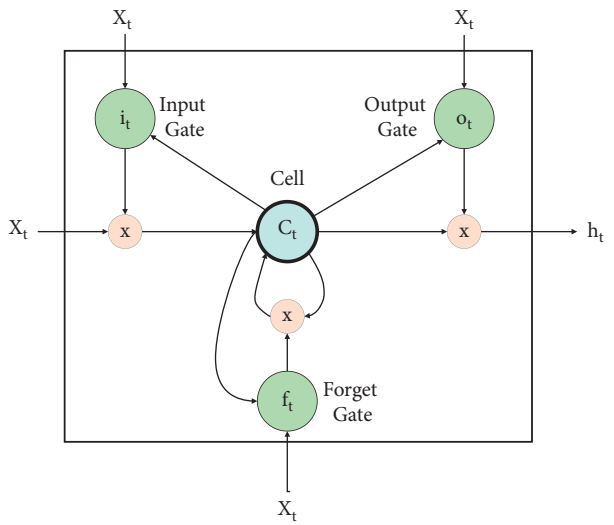

Figure 6: Long short-term memory (LSTM).

8.1. Number of New FinTech Patents. As shown in Figure 9, the number of FinTech patent applications in China was 862 in 2013 and increased each year from 2013 to 2018. Between 2018 and 2020, the number of patents remained high and stable, but the shares of different patent types changed significantly, reflecting the vigorous development of FinTech innovation. The growth in the number of FinTech patent applications in the payment category peaked (1029) in 2018 and then slowed. This shift may reflect the increasing maturity of mobile payment technology and, as a result, limited room for innovation. At present, the main innovations in payment FinTech are improving existing mobile payment systems and making mobile payments more convenient and faster by using Internet of Things technology. The number of data analytics FinTech patent applications rose each year, heralding the era of big data and cloud computing technology and financial industry integration. The future directions of development of financial science and technology will be efficiently handling large databases, upgrading financial service models, and promoting the development of the financial industry.

Figure 10 shows regional differences in the number of FinTech patent applications in China. FinTech-related patent applications from Guangdong and Beijing accounted for 52.47% of all FinTech patent applications in China. Overall, the number of FinTech patent applications was higher in economically developed regions than in less-developed regions. Provinces with more than 1000 applications included Guangdong, Beijing, Shanghai, Jiangsu, and Jiangsu. Figure 11 shows the proportions of the annual number of FinTech patent applications in these five provinces compared with all other provinces.

Comparing the proportions of FinTech patent applications in various provinces from 2013 to 2020 provides insights into the development process of FinTech in China. FinTech innovation gradually concentrated in the developed provinces. The proportions of FinTech patent applications from Beijing, Shanghai, Guangzhou, Jiangsu, and Zhejiang increased over the analyzed period, and in 2020, Beijing surpassed Guangzhou to lead China in the number of FinTech patent applications. Along with this growth, the proportion of payment patents decreased, whereas the proportion of data analytics patents increased, as shown in Figure 11. Accordingly, we can speculate that the change in patent application structure at the provincial level reflects the shift in the focus of FinTech development to the analysis and prediction of user behavior from big data. Beijing and Guangdong have advantages over other

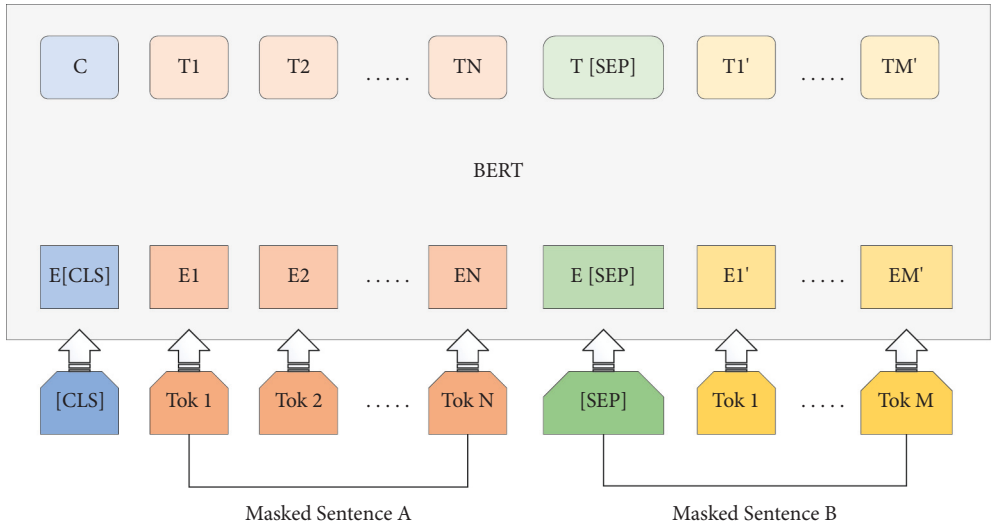

Figure 7: Pretraining BERT model structure.

Table 5: The classification performance of the different models on the test set.

Category	Metric	SVM	KNN	NB	DT	CNN	RNN	LSTM	BERT
FinTech-unrelated	Pr	0.74	0.54	0.47	0.62	0.81	0.47	0.44	0.84*
	Re	0.67	0.76	0.34	0.6	0.87	0.32	0.56	0.9*
	F1	0.71	0.63	0.40	0.61	0.84	0.38	0.50	0.87*
Insurance	Pr	0.58	0.45	0.46	0.34	0.82*	0.42	0.37	0.81
	Re	0.30	0.56	0.22	0.32	0.90	0.30	0.34	0.95*
	F1	0.39	0.50	0.30	0.33	0.86	0.35	0.35	0.87*
Data analytics	Pr	0.57	0.63	0.67	0.63	0.78*	0.67	0.65	0.74
	Re	0.73	0.76*	0.54	0.71	0.58	0.54	0.59	0.55
	F1	0.65	0.69*	0.59	0.67	0.67	0.59	0.62	0.63
Lending	Pr	0.92*	0.62	0.85	0.80	0.84	0.62	0.87	0.86
	Re	0.68	0.72	0.44	0.74	0.98*	0.42	0.66	0.96
	F1	0.78	0.67	0.58	0.77	0.91*	0.50	0.75	0.91*
Payment	Pr	0.76	0.62	0.40	0.59	0.90*	0.62	0.66	0.90*
	Re	0.74	0.78	0.24	0.70	0.86*	0.42	0.61	0.78
	F1	0.75	0.69	0.30	0.64	0.88*	0.50	0.63	0.83
Investment	Pr	0.90*	0.84	0.78	0.82	0.82	0.88	0.76	0.77
	Re	0.72	0.64	0.64	0.74	0.82*	0.60	0.74	0.82*
	F1	0.80	0.73	0.70	0.78	0.82*	0.71	0.75	0.8
Security	Pr	0.67	0.46	0.42	0.45	0.89*	0.44	0.62	0.86
	Re	0.67	0.59	0.33	0.47	0.80*	0.28	0.49	0.66
	F1	0.67	0.52	0.37	0.46	0.85*	0.35	0.55	0.75
Average	Pr	0.72	0.64	0.55	0.63	0.84*	0.59	0.65	0.83
	Re	0.72	0.61	0.54	0.63	0.83*	0.57	0.64	0.80
	F1	0.71	0.61	0.52	0.63	0.83*	0.55	0.64	0.81
	Acc	0.72	0.61	0.53	0.63	0.84*	0.56	0.64	0.83

The significance of the symbol "*" means that the model performs significantly better than others in the patent classification.

provinces in both development environment and technology, and the development of big data and cloud computing technology is relatively mature. Consequently, it is easy for these provinces to master this technology and apply it to the financial industry to lead the development of FinTech.

Figure 12 presents the numbers of FinTech patent applications by financial and nonfinancial sectors. In the financial sectors, banks are the main drivers of FinTech innovation due to the size and complexity of their financial operations. In 2019 and 2020, the number of FinTech patent applications filed by banks, which had few applications compared with nonfinancial sectors before 2019, began to rise rapidly. Prior to 2019, most FinTech patent applications came from Bank of China Limited and Shenzhen Qianhai WeBank Limited. Although an important part of the financial industry, banks were initially not sensitive to the development of financial technology and had a low level of

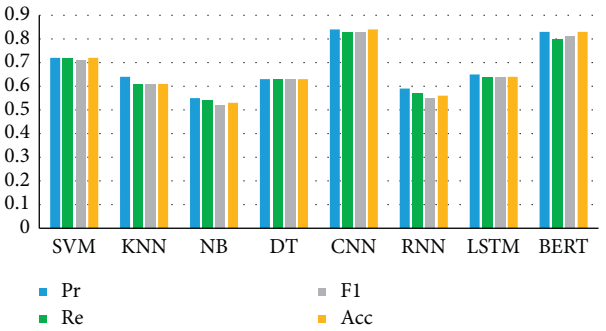

Figure 8: Average metrics of the different models.

Table 6: Complete FinTech patent classification process.

Filtering and sorting steps	Number of remaining samples	Number of remaining samples
All patents under entries G and H filed in China from 2013 to 2020 in the database		5398266
Remove pure technical patent data from IPC classification	4072571	1325745
Filter nonfinancial patents via machine learning	1248327	77418
Classify text using CNN to remove FinTech-unrelated patents	52265	25153
Remove patents filed abroad	4624	20529
Number of remaining FinTech patent applications	20529	
FinTech patents		
Individual patent applications	1967	
Nonindividual patent applications	18562	

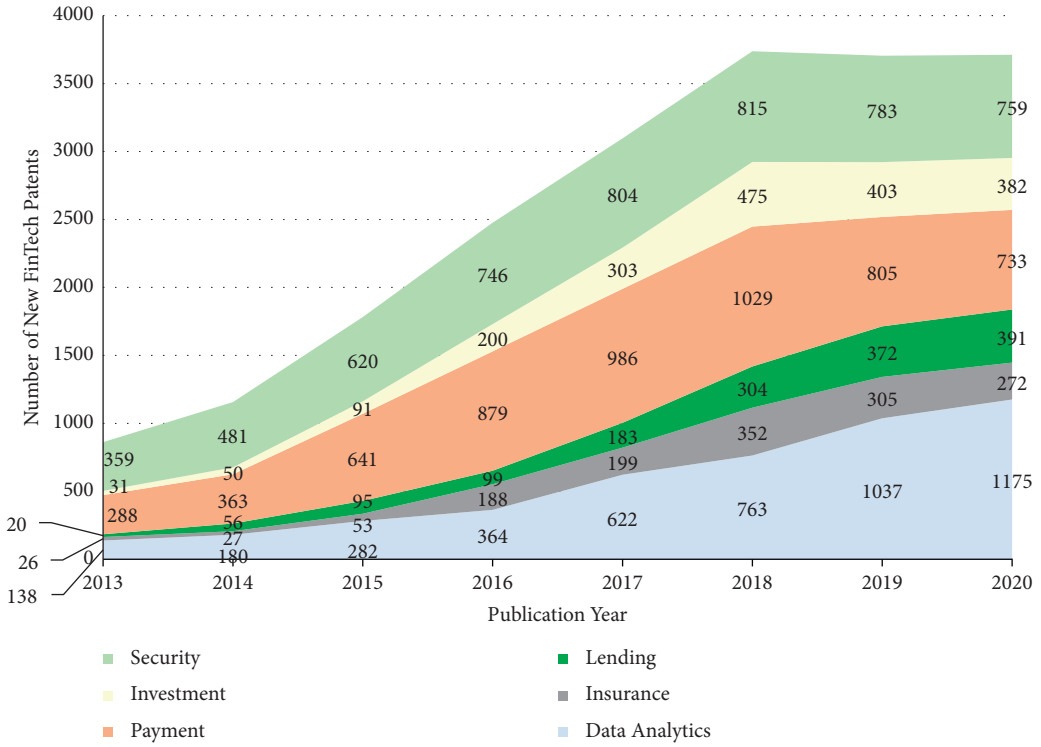

Figure 9: Annual number of new FinTech patent applications in China.

innovation. To some extent, this explains why Alipay and WeChat, rather than banks, currently dominate China's mobile payment business. In recent years, the traditional banking business has been strongly impacted by the development of FinTech. Various online lending institutions have entered traditional banking business activities such as

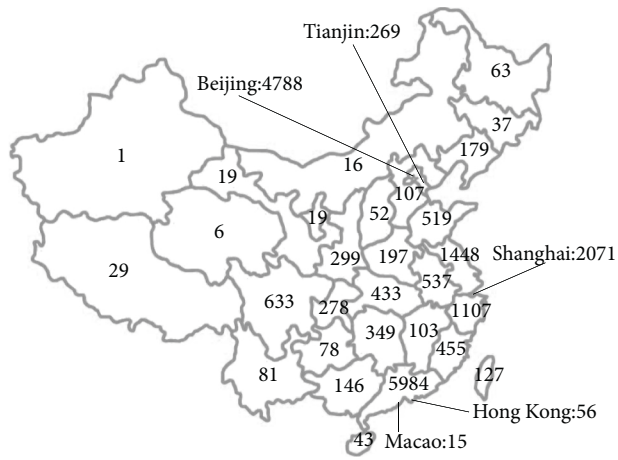

Figure 10: The distribution of FinTech patent applications at the provincial level in China.

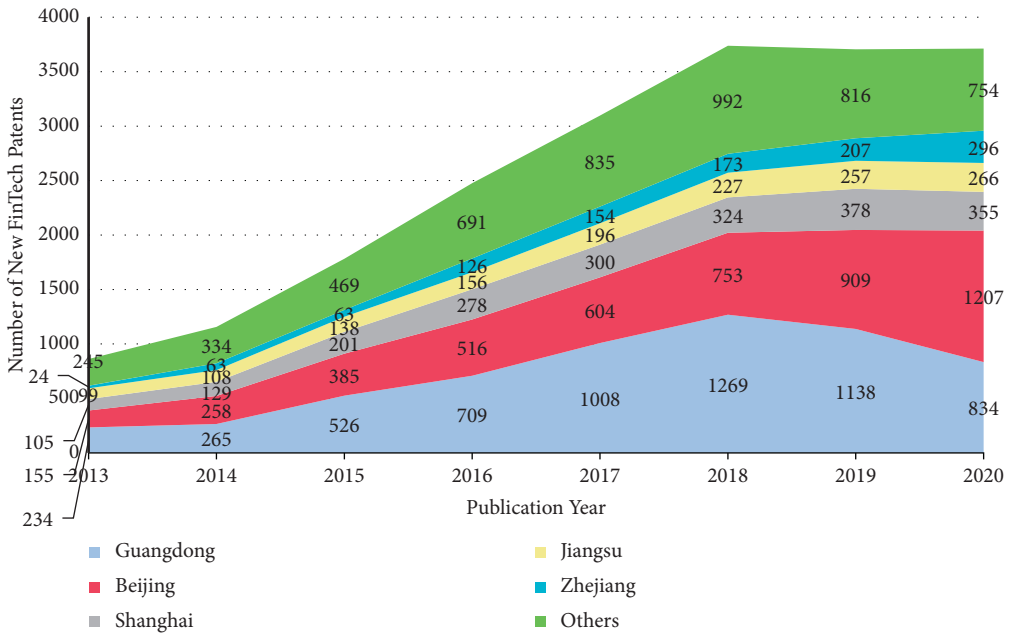

Figure 11: Proportions of the annual number of FinTech patent applications by province in China.

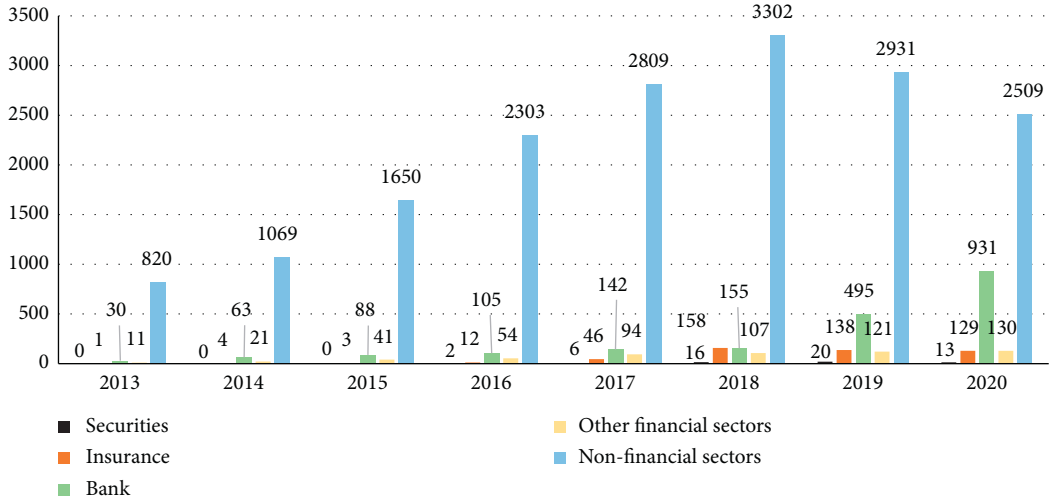

Figure 12: Annual number of FinTech patent applications in the financial and nonfinancial sectors.

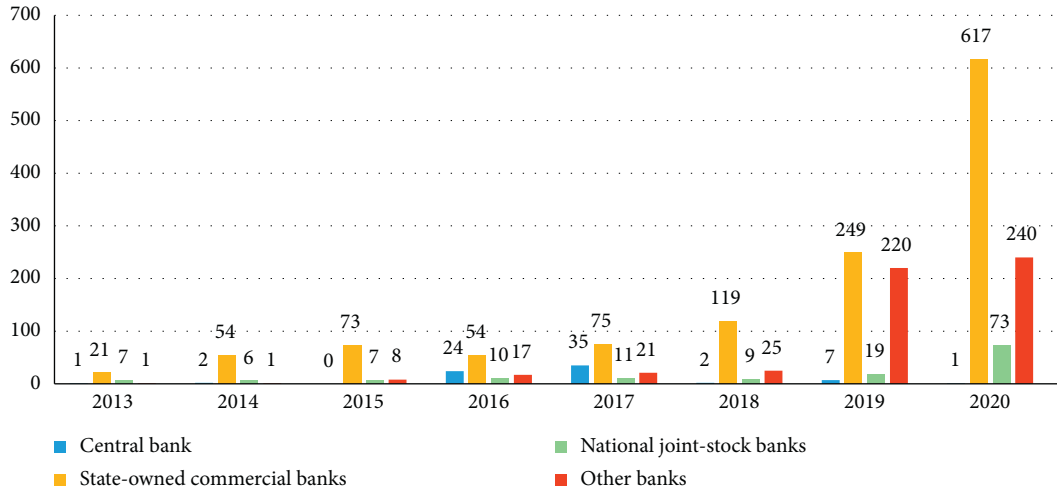

Figure 13: Annual number of FinTech patent applications of banks.

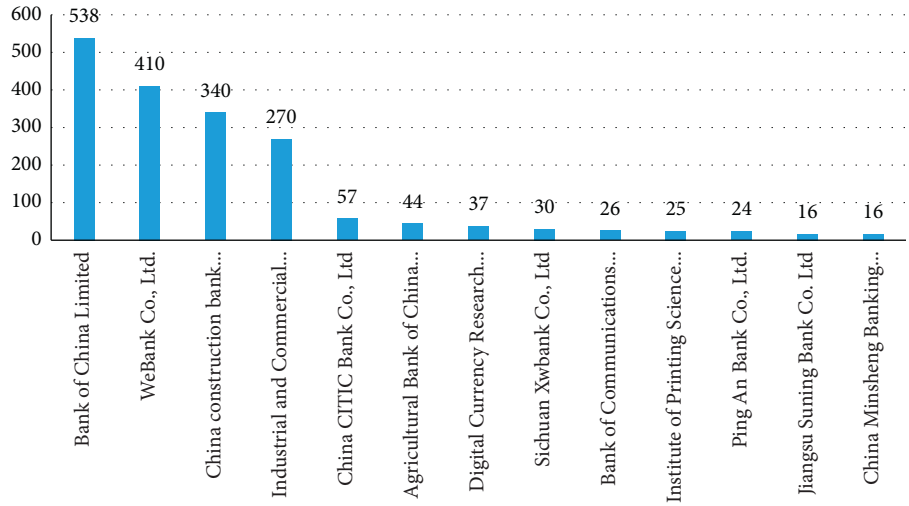

Figure 14: Banks with the highest numbers of FinTech patents.

lending. In response, banks have intensified their FinTech innovation efforts. The next round of banking competition may center on FinTech innovation. Relevant research on banking competition suggests that this industry trend may impact the development of local enterprises and the allocation of credit resources.

In China, the banking industry can be divided into central bank, state-owned commercial banks, national joint-stock commercial banks, and other banks. Figure 13 shows the number of FinTech patent applications per year for each type of bank. The central bank has a certain number of FinTech patent applications between 2016 and 2017, most of which are from the digital currency research institute of the central bank, which are products of China's digital currency strategy. The FinTech innovation of large state-owned banks has developed rapidly in recent years, which may be driven by their huge financial business and the need for digital transformation under internal and external pressures. What is more, from Figure 14, we can see that, among other banks, WeBank accounted for the lion's share of FinTech patent applications, and FinTech innovation has a high concentration.

9. Conclusions and Future Work

Based on an analysis of the connotation and function of FinTech as well as prior research related to FinTech, this study divided Chinese FinTech patents into six categories: payment, lending, insurance, security, data analysis, and investment. Using data from all patent applications published by China's State Intellectual Property Office between 2013 and 2020, we selected and classified FinTech-related patents using traditional machine learning and deep learning approaches. First, we used text filtering and manual annotation to obtain an annotated dataset for model training; then, the training and test datasets were used with different machine learning and deep learning models to compare the models' effectiveness in text categorization.

Character-level CNN gave the best classification performance and was selected to classify the patent data, ultimately resulting in the identification of 20,529 FinTech-related patents.

The classified data were then used to discuss the current situation of FinTech innovation in China. The number of annual patent applications for payment FinTech reached its peak in 2018. The high intensity of innovation led to prosperity in the mobile payment business, in line with the development of payment systems and the huge scale of business in China. Currently, data analytics patent applications account for the largest share of FinTech patent applications, indicating the future direction of FinTech.

There are distinct regional differences in the number of FinTech patent applications in China, with more applications in developed provinces than in less-developed provinces. Moreover, the concentration of new FinTech patent applications in Beijing, Shanghai, and Guangzhou is increasing each year. At the microlevel, China's banks entered FinTech innovation relatively late. Since 2019, the traditional banking industry has caught up with the FinTech challenge, and the number of FinTech patent applications from banks has grown rapidly.

The main contribution of this study is that we provide a way to search for Chinese FinTech patents in order to research on FinTech innovation while also helping the financial firms and regulators to understand the latest developments in FinTech. This study is the first study on the classification of Chinese FinTech patents. The dataset we constructed can deepen the research on FinTech innovation in China from the macrolevel to the microlevel; to be specific, it could be used to study where, when, and why FinTech innovations emerge and what their impact is on firms investing in them.

In future research, we plan to further associate FinTech patent application data with corporate microdata, industry mesodata, and economic macrodata to more deeply study the impact of FinTech innovation on the allocation of financial resources and corporate governance in China.

Data Availability

The data used to support the findings of this study are included within the supplementary information file.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC no. 72103083).

Supplementary Materials

The dataset contains all FinTech patent information, including patent name, applicant, application date, public date, number, address, and classification. (*Supplementary Materials*)

References

[1] I. Goldstein and W. Jiang, "To FinTech and beyond," *Review of Financial Studies*, vol. 32, no. 5, pp. 1647–1661, 2019.

[2] M. Harrist, "How FinTech is powering the global economy," Forbes, 2017.

[3] A. Fuster, M. Plosser, P. Schnabl, and J. Vickery, "The role of technology in mortgage lending," *Review of Financial Studies*, vol. 32, no. 5, pp. 1854–1899, 2019.

[4] H. Tang, "Peer-to-peer lenders versus banks: substitutes or complements?" *Review of Financial Studies*, vol. 32, no. 5, pp. 1900–1938, 2019.

[5] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the rise of FinTechs: credit scoring using digital footprints," *Review of Financial Studies*, vol. 33, no. 7, pp. 2845–2897, 2020.

[6] C. Haddad and L. Hornuf, "The emergence of the global FinTech market: economic and technological determinants," *Small Business Economics*, vol. 53, no. 1, pp. 81–105, 2019.

[7] P. Utami and B. Basrowi, "Management of zakat payment based on FinTech for the good corporate governance improvement," *Eastern Journal of Economics and Finance*, vol. 4, no. 2, pp. 41–50, 2019.

[8] M. A. Chen, Q. Wu, and B. Yang, "How valuable is FinTech innovation?" *Review of Financial Studies*, vol. 32, no. 5, pp. 2062–2106, 2019.

[9] L. Xu, X. Lu, G. Yang, and B. Shi, "Identifying FinTech innovations with patent data: a combination of textual analysis and machine-learning techniques," in *Proceedings of the International Conference on Information*, pp. 835–843, Springer, Shanghai, China, September 2020.

[10] D. Caragea, M. Chen, T. Cojoianu, M. Dobri, K. Glandt, and M. George, "Identifying FinTech innovations using BERT," in *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pp. 1117–1126, IEEE, Atlanta, Georgia, US, December 2020.

[11] J. Xu, "China's internet finance: a critical review," *China and World Economy*, vol. 25, pp. 78–92, 2017.

[12] Kamdjoug, R. E. Bawack, and J. G. Keogh, "Bitcoin, Blockchain and FinTech: a systematic review and case studies in the supply chain," *Production Planning & Control*, vol. 31, no. 2-3, pp. 115–142, 2020.

[13] T. Nakashima, "Creating credit by making use of mobility with FinTech and IoT," *IATSS Research*, vol. 42, no. 2, pp. 61–66, 2018.

[14] W. Du, S. L. Pan, D. E. Leidner, and W. Ying, "Affordances, experimentation and actualization of FinTech: a blockchain implementation study," *The Journal of Strategic Information Systems*, vol. 28, no. 1, pp. 50–65, 2019.

[15] I. Romānova and M. Kudinska, "Banking and FinTech: a challenge or opportunity?" in *Contemporary Issues in Finance: Current Challenges from across Europe*Emerald Group Publishing Limited, Bingley, United Kingdom, 2016.

[16] Y. J. Shin, "FinTech: ecosystem, business models, investment decisions, and challenges," *Business Horizons*, vol. 61, no. 1, pp. 35–46, 2018.

[17] M. Demertzis, S. Merler, and G. B. Wolff, "Capital markets union and the FinTech opportunity," *Journal of financial regulation*, vol. 4, no. 1, pp. 157–165, 2018.

[18] G. Buchak, P. Matvos, and A. Seru, "FinTech, regulatory arbitrage, and the rise of shadow banks," *Journal of Financial Economics*, vol. 130, no. 3, pp. 453–483, 2018.

[19] A. V. Thakor, "FinTech and banking: what do we know?" *Journal of Financial Intermediation*, vol. 41, Article ID 100833, 2020.

[20] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, "Automated categorization in the international patent classification," *Acm Sigir Forum*, vol. 37, no. 1, pp. 10–25, 2003.
[21] K. Benzineb and J. Guyot, "Automated patent classification, current challenges in patent information retrieval," in *Current Challenges in Patent Information Retrieval*, pp. 239–261, Springer, Berlin, Heidelberg, 2011.
[22] L. Aristodemou and F. Tietze, "The state-of-the-art on Intellectual Property Analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data," *World Patent Information*, vol. 55, pp. 37–51, 2018.
[23] C.-H. Wu, Y. Ken, and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine," *Applied Soft Computing*, vol. 10, no. 4, pp. 1164–1177, 2010.
[24] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, 2018.
[25] L. Xiao, G. Wang, and Z. Yang, "Research on patent text classification based on word2vec and LSTM," in *Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 71–74, IEEE, Hangzhou, China, December 2018.
[26] J. Zhao, X. Li, C.-H. Yu, S. Chen, and C.-C. Lee, "Riding the FinTech innovation wave: FinTech, patents and bank performance," *Journal of International Money and Finance*, vol. 122, no. 2022, Article ID 102552, 2022.
[27] T. F. Cojoianu, G. L. Clark, A. G. F Hoepner, V Pažitka, and D. Wójcik, "Fin vs. tech: are trust and knowledge creation key ingredients in fintech start-up emergence and financing?" *Small Business Economics*, vol. 57, no. 4, pp. 1715–1731, 2021.
[28] W. Lee and S. Sohn, "Identifying emerging trends of financial business method patents," *Sustainability*, vol. 9, no. 9, p. 1670, 2017.
[29] L. Chen, "From FinTech to finlife: the case of FinTech development in China," *China Economic Journal*, vol. 9, no. 3, pp. 225–239, 2016.
[30] Y. Shim and D.-H. Shin, "Analyzing China's fintech industry from the perspective of actor-network theory," *Telecommunications Policy*, vol. 40, no. 2-3, pp. 168–181, 2016.
[31] C. Leong, B. Tan, F. T. C. Tan, and Y. Sun, "Nurturing a FinTech ecosystem: the case of a youth microloan startup in China," *International Journal of Information Management*, vol. 37, no. 2, pp. 92–97, 2017.