Financial Risk Prediction and Management using Machine Learning and Natural Language Processing

Tianyu Li, Xiangyu Dai
Hunan Vocational College of Commerce, Changsha, China

***Abstract*—With the continuous development and changes in the global financial markets, financial risk management has become increasingly important for the stable operation of enterprises. Traditional financial risk management methods, primarily relying on financial statement analysis and historical data statistics, show clear limitations when dealing with large-scale unstructured data. The rapid development of machine learning and Natural Language Processing (NLP) technologies in recent years offers new perspectives and methods for financial risk prediction and management. This paper explores and conducts empirical analysis financial risk management using these advanced technologies, with a particular focus on the application of NLP in measuring financial risk tendencies, and the financial risk prediction and management based on a Deep neural network - Factorization Machine (DeepFM) model. Through in-depth analysis and research, this paper proposes a new financial risk management model that combines NLP and deep learning technologies, aimed at improving the accuracy and efficiency of financial risk prediction. This study not only broadens the theoretical horizons of financial risk management but also provides effective technical support and decision-making references for practical operations.**

Keywords—Financial risk management; machine learning; Natural Language Processing (NLP); Deep FM model; risk prediction

I. Introduction

In today’s rapidly developing financial industry, financial risk management has become crucial for the survival and development of enterprises [1-3]. With the rapid progress of big data technology and machine learning, how to effectively use these advanced technologies to predict and manage financial risks has become a hot topic of research and practice [4, 5]. Traditional financial risk management methods often rely on financial statement analysis and historical data statistics, but they show clear limitations in dealing with large-scale unstructured data, such as news texts and social media information [6-8]. Therefore, exploring a new method of financial risk prediction and management is particularly important.

In recent years, the application of machine learning and NLP technologies in the financial field has become increasingly widespread, especially showing great potential in financial risk prediction and management [9-11]. By analyzing a large amount of historical data and real-time information, these technologies can not only identify and evaluate potential financial risks but also provide more accurate predictions, helping enterprises to make more rational decisions. However, how to effectively integrate these technologies and apply them to financial risk management, as well as how to process and analyze large-scale unstructured data, remains a question that requires in-depth research [12-14].

Although current research on the application of machine learning and NLP in financial risk management is gradually increasing, most studies focus on the application of specific models and lack an in-depth discussion on the integrated application of different technologies [15, 16]. Moreover, existing research still has deficiencies in dealing with unstructured data, especially in the application of deep understanding and sentiment analysis of text data, which limits its accuracy and effectiveness in financial risk prediction [17-20].

This paper aims to explore the methods of big data financial risk prediction and management based on machine learning and NLP. Firstly, this study measures financial risk tendencies through NLP technology, effectively extracting and analyzing unstructured text data from various channels, providing a richer dimension for risk assessment. Secondly, this paper introduces a financial risk prediction model based on Deep FM, which can effectively integrate various features and improve the accuracy and efficiency of predictions through deep learning technology. Through these two aspects of research, this paper not only expands the theory and methods of financial risk management but also provides new ideas and tools for practical application, having significant theoretical significance and practical value.

II. Measurement of Financial Risk Propensity Using NLP

In financial risk management, NLP technologies are employed to analyze and quantify the propensity of financial risks. These technologies extract and process key information from unstructured textual data across various channels, such as news articles, financial reports, and social media feeds, offering a more comprehensive perspective on risk assessment. Subsequently, a financial risk prediction model based on the DeepFM algorithm integrates these insights derived from textual data with a multitude of other features, leveraging the power of deep learning to enhance the accuracy and efficiency of risk forecasting. The ultimate goal is to achieve more precise and effective financial risk management. This chapter discusses the specific implementation details of the financial risk propensity measurement model based on the CSBL algorithm. For a corpus containing V financial-related comments, each comment consisting of J words, let $A=\{A_1,A_2,...,A_V\}$ be a specific comment in the corpus, $A_v=\{A_1,A_2,...,A_j\}$ represents a set of vocabulary in the comment

A_v, each vocabulary A_j is an F-dimensional embedding word vector. The model aims to predict the financial risk tendency B for each comment, where B only includes the sentiment tendency of financial risk rising or falling. To ensure the clarity of the input data's sentiment tendency, this model specifically filters out those comments that express neutral, objective, or unclear financial viewpoints. The financial risk characteristics of each comment are represented by discrete values and encoded using the *one-hot* encoding method, mapping the financial risk characteristics of the comments to two-bit *one-hot* encoding: [0, 1] represents financial risk rising (negative emotion), [1, 0] indicates financial risk falling (positive emotion). In this way, we can obtain the financial risk propensity label $Yn(Yn \in \{[0, 1], [1, 0]\})$ for any comment Xn, thereby accurately measuring and predicting the risk tendency of financial texts.

In the measurement of financial risk propensity using NLP, the application of the CSBL model involves several key steps aimed at accurately extracting and analyzing risk information from financial texts. First, the model thoroughly preprocesses input texts such as financial reports, press releases, or market comments to optimize subsequent feature extraction and risk propensity analysis. Next, by constructing a specialized capsule network, the model adjusts and highlights important financial risk features. Subsequently, it uses a Stacked Bi-LSTM network to delve into the contextual relationships of the texts, thereby capturing potential risk signals. Finally, the model inputs the comprehensive representation of these complex features into a softmax classifier to predict the financial text's risk propensity, i.e., risk rising or falling. Fig. 1 shows the architecture of the Stacked-BiLSTM network model.

1) In the preprocessing stage, the model first determines the average financial text length J, which serves as the standard size for network input, and normalizes the length of input texts accordingly. For texts exceeding J words, the exceeding part is truncated; for those less than J words, zero-padding is used to reach the length of J words. Additionally, each vocabulary is mapped to an F-dimensional embedding vector, and vocabularies not found in the word embedding model are replaced with F-dimensional zero vectors. This series of preprocessing steps ensures that each financial text is converted into a uniform J*F-dimensional vector format, providing the model with standardized and information-rich input, thereby laying a solid foundation for subsequent risk propensity analysis. Through this refined preprocessing process, the model can effectively process and analyze various financial texts, providing more accurate and comprehensive support for financial risk prediction and management.

2) In the financial risk propensity measurement model based on NLP, the second step's key is using the capsule network to adjust the weight of important features in financial texts. Unlike traditional neural networks with scalar neuron nodes, neurons in the capsule network exist in vector form, allowing the network to strengthen the representation weight of important features through a dynamic routing algorithm during training. This algorithm optimizes the feature selection process, enabling the model to reveal more hidden financial risk-related features, thereby significantly enhancing the model's performance in financial risk propensity analysis. In this stage, the model adopts the word2vec method to convert texts into vector form, which are then input into the capsule network for further feature extraction and weighting.

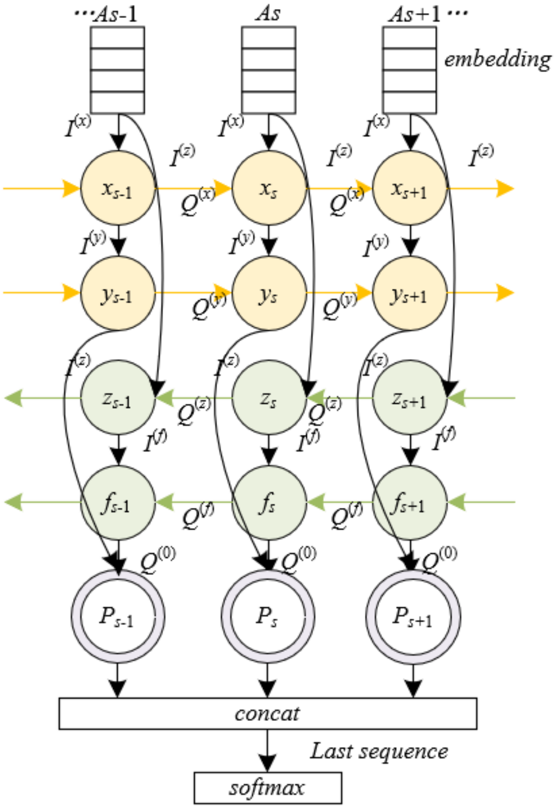

Fig. 1. Architecture of the stacked-BiLSTM network model.

Particularly, the processing of financial texts takes into account the unique structure and hierarchy of the text. Convolutional layers in the capsule network extract n-gram features from sentences through convolutional filters, slightly different from image processing methods. The text input a is represented as an M (sentence length) * N (embedding word dimension) matrix, where each a_u represents an N-dimensional word vector. Convolutional filters Q^{β}, with a length of M-$J1$+1, slide through the sentence in an n-gram manner ($J1$ being the n-gram length, i.e., the size of the sliding window), capturing text features at different positions. Each time the filter slides to a new position, it generates a feature map l^x, these mappings, after going through the dynamic routing process of the capsule network, effectively highlight the key text features related to financial risk. This process not only enhances the model's sensitivity to financial risk propensity analysis but also provides high-quality feature representations for subsequent steps, laying a solid foundation for accurately predicting the risk propensity of financial texts. Let unit multiplication be represented by p, bias by y_0, the nonlinear activation function by d, and the sliding stride by m. Then, the expression for l^x is as follows:

$$\mathrm{x} \quad l_m^x = d\left(a_{u:u+j_1-m} \circ Q^{\beta} + y_0\right) \tag{1}$$

For Y filters with the same n-gram size, the following Y-dimensional feature mapping can be generated and reordered:

$$L = [l_1, l_2, \ldots, l_Y] \tag{2}$$

In the application of the capsule network for measuring financial risk propensity, the design of the capsule layer allows the model to retain more information when processing financial text data. Fig. 2 shows the architecture of the capsule network. Traditional neural networks use scalar outputs to represent the activation state of neurons, whereas capsule networks employ vector outputs. This is done to preserve instantiation parameters within the text data, such as context and word order, which are crucial for understanding the complexity of financial texts. Specifically, the output neurons L generated by the convolutional layer serve as the input vectors for the capsule layer. By applying the activation function, the model converts each n-gram feature vector L_u into its corresponding feature capsule i_u. This step further transforms the extracted text features into capsule vectors capable of representing financial risk information, laying the foundation for the subsequent dynamic routing process. Assuming that the filters shared by different sliding windows are represented by Q_y, the capsule bias by y_1, the nonlinear activation function by h, and the weight matrix of the correlation between the input and output layers by Q_{uk}, the formulas for converting L_u into i_u using the activation function are:

$$I_u = h(Q_y L_u + y_1) \tag{3}$$

$$i_{k/u} = Q_{uk} I_u \tag{4}$$

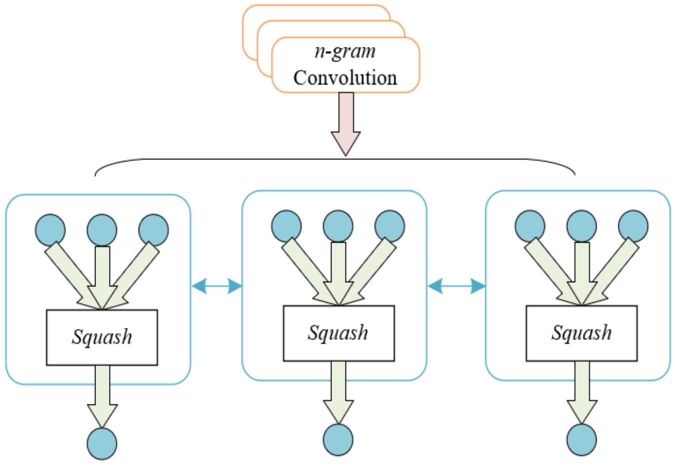

Fig. 2. Capsule network architecture.

One of the core aspects of the capsule network is its dynamic routing process, which optimizes the information transfer between feature capsules by updating the weights of the coupling coefficients. The update of coupling coefficients depends on the similarity between adjacent capsules; that is, the more similar the output vectors of two capsules, the greater their coupling coefficient. This similarity-based dynamic routing strategy is not only more efficient than the routing mechanism in traditional Convolutional Neural Networks (CNNs) but also ensures that capsules carrying significant financial risk signals are accurately passed to the next layer of the network. In measuring financial risk propensity, this means the model can more accurately identify and reinforce those text features that are crucial for predicting financial risk propensity.

The execution of dynamic routing involves considering each predicted vector $i_{k/u}$ and its existence probability $x_{k/u}$, optimizing the feature selection and information transfer paths of the entire network by iteratively updating the coupling coefficients z_{uk}. The intent of this process is to enhance the representation of similarity between input vectors and the target classification, assigning higher weights to those capsule outputs n_k and predicted vectors $i_{k/u}$ that are closer to each other. The initial value of the coupling coefficients y_{uk} is set to 0, and through the iterative process, the model adaptively adjusts these coefficients to ultimately achieve accurate prediction of financial risk propensity. Additionally, the dynamic routing includes a squashing function that ensures the absolute value of the input vectors is compressed into the range [0,1), further increasing the model's flexibility and accuracy in processing financial text data. Through this series of refined processes, the capsule network provides strong technical support for efficient and accurate measurement of financial risk propensity. Below are the expressions for the dynamic routing process:

$$z_{uk} = x_{k/u} \cdot \exp(y_{uk}) \frac{\exp(y_{uk})}{\exp(y_{uj})} \tag{5}$$

$$T_k = \sum_k z_{uk} i_{k/u} \tag{6}$$

$$n_k = \frac{\| T_k \|^2}{1 + \| T_k \|^2} \times \frac{T_k}{\| T_k \|}, x_k = | N_k | \tag{7}$$

$$y_{uk} \leftarrow y_{uk} + i_{k/u} \cdot n_k \tag{8}$$

3) Extracting contextual features of documents is a key step in the implementation of the financial risk propensity measurement model, accomplished through the use of a Stacked-BiLSTM network. Compared to the standard BiLSTM network, Stacked-BiLSTM has multiple hidden layers, enabling the model to perform deeper feature extraction. By setting multiple layers of LSTM both forward and backward in time series, Stacked-BiLSTM can capture both past and future contextual information, providing a richer and more detailed feature representation for accurate prediction of financial risk propensity. This capability is particularly important in financial text analysis, as risk signals in documents like financial reports and market comments are often closely related to a complex context, requiring the model to consider temporal features and contextual dependencies within the text comprehensively.

Regarding the structure of the Stacked-BiLSTM network, the input sequence at each time point $\{a_1, a_2, \ldots, a_S\}$ is processed by multiple layers of LSTM in both forward and backward directions to capture more feature information from each time step. Each LSTM layer contains new memory cells, input gates, forget gates, and output gates, represented by i_s, u_s, d_s, and p_s, respectively. These components collectively decide how to update states, store, or forget information, and determine which information will be passed to the next layer of the network. This design allows the Stacked-BiLSTM network to effectively control the flow of information when processing financial texts, retaining the most critical features for risk prediction while ignoring irrelevant or redundant information. Specifically, assuming $\{a_1,a_2,\ldots,a_S\}$ enters the hidden layer in the forward direction $\{x_1,x_2,\ldots,x_S\}$, and captures more features from all subsequent time steps in the opposite direction's hidden layer $\{z_1,z_2,\ldots,z_S\}$. The hidden state of each layer at every time step s is represented by x_s, y_s, z_s, and f_s.

The following gives the calculation formula for the hidden state x_s of the first forward layer:

$$\begin{cases} u_s{}^{(x)} = \delta\left(I_u^{(x)} a_s + Q_u^{(x)} x_{s-1} y_u^{(x)}\right), \\ d_s^{(x)} = \delta\left(I_d^{(x)} a_s + Q_d^{(x)} x_{s-1} + y_d^{(x)}\right), \\ p_s^{(x)} = \delta\left(I_p^{(x)} a_s + Q_p^{(x)} x_{s-1} + y_p^{(x)}\right), \\ i_s^{(x)} = TANg\left(I_i^{(x)} a_s + Q_i^{(x)} x_{s-1} + y_i^{(x)}\right), \\ Z_s^{(x)} = u_s^{(x)} * i_s^{(x)} + d_s^{(x)} * Z_{s-1}^{(x)}, \\ x_s = p_s^{(x)} * TANg\left(Z_s^{(x)}\right). \end{cases} \tag{9}$$

The formula for calculating the hidden state y_s of the second forward layer is:

$$\begin{cases} u_s{}^{(y)} = \delta\left(I_u^{(y)} a_s + Q_u^{(y)} x_{s-1} y_u^{(y)}\right), \\ d_s^{(y)} = \delta\left(I_d^{(y)} a_s + Q_d^{(y)} x_{s-1} + y_d^{(y)}\right), \\ p_s^{(y)} = \delta\left(I_p^{(y)} a_s + Q_p^{(y)} x_{s-1} + y_p^{(y)}\right), \\ i_s^{(y)} = TANg\left(I_i^{(y)} a_s + Q_i^{(y)} x_{s-1} + y_i^{(y)}\right), \\ Z_s^{(y)} = u_s^{(.y)} * i_s^{(y)} + d_s^{(y)} * Z_{s-1}^{(y)}, \\ x_s = p_s^{(y)} * TANg\left(Z_s^{(y)}\right). \end{cases} \tag{10}$$

The formula for calculating the hidden state z_s of the first backward layer is:

$$\begin{cases} u_s{}^{(z)} = \delta\left(I_u^{(z)} a_s + Q_u^{(z)} x_{s+1} y_u^{(z)}\right), \\ d_s^{(z)} = \delta\left(I_d^{(z)} a_s + Q_d^{(z)} x_{s+1} + y_d^{(z)}\right), \\ p_s^{(z)} = \delta\left(I_p^{(z)} a_s + Q_p^{(z)} x_{s+1} + y_p^{(z)}\right), \\ i_s^{(z)} = TANg\left(I_i^{(z)} a_s + Q_i^{(z)} x_{s+1} + y_i^{(z)}\right), \\ Z_s^{(z)} = u_s^{(z)} * i_s^{(z)} + d_s^{(z)} * Z_{s-1}^{(z)}, \\ x_s = p_s^{(z)} * TANg\left(Z_s^{(z)}\right). \end{cases} \tag{11}$$

The formula for calculating the hidden state of the second backward layer is:

$$\begin{cases} u_s{}^{(f)} = \delta\left(I_u^{(f)} a_s + Q_u^{(f)} x_{s+1} y_u^{(f)}\right), \\ d_s^{(f)} = \delta\left(I_d^{(f)} a_s + Q_d^{(f)} x_{s+1} + y_d^{(f)}\right), \\ p_s^{(f)} = \delta\left(I_p^{(f)} a_s + Q_p^{(f)} x_{s+1} + y_p^{(f)}\right), \\ i_s^{(f)} = TANg\left(I_i^{(f)} a_s + Q_i^{(f)} x_{s+1} + y_i^{(f)}\right), \\ Z_s^{(f)} = u_s^{(f)} * i_s^{(f)} + d_s^{(f)} * Z_{s-1}^{(f)}, \\ x_s = p_s^{(f)} * TANg\left(Z_s^{(f)}\right). \end{cases} \tag{12}$$

For each time step s, the output P_s is generated by combining y_s and f_s, as follows:

$$P_s = I^{(P)} y_s + Q^{(P)} f_s + y^{(P)} \tag{13}$$

To output financial risk propensity prediction results, the Softmax classifier takes P_j as its input. Given V comments and J words, the prediction value b' is calculated as follows:

$$o\left(b \mid A\right) = \text{softmax}\left(Q^{(t)} P_J + y^{(t)}\right) \tag{14}$$

$$b' = \underset{b}{\text{argmax}}\, o\left(b \mid A\right). \tag{15}$$

III. Financial Risk Prediction and Management Based on Deep FM in Big Data

Financial data typically includes but is not limited to, transaction records, financial statements, market dynamics, etc., which contain complex feature relationships, including both linear and nonlinear interactions, posing a challenge to traditional prediction models. This paper introduces the DeepFM model. By integrating factorization machines and deep neural networks, DeepFM can capture not only the linear relationships between features but also learn higher-order feature combinations, which is crucial for understanding and predicting financial risks. Compared to other fields, financial risk management demands higher accuracy and efficiency in predictions, and the DeepFM model meets these dual requirements of model performance and efficiency by sharing feature embedding vectors, reducing the model's parameter amount and computational cost, while ensuring fast training and prediction speeds.

The construction process of the DeepFM model in the application of big data financial risk prediction and management includes three key stages: feature combination, efficient feature representation, and classification prediction. Firstly, the input features are combined through the FM part, utilizing FM's advantage to capture the interactions between features. This step is particularly suited for handling the rich low-order feature interactions in financial data, providing a foundation for capturing complex financial risk patterns. Subsequently, in the Deep Neural Network (Deep) part, the model uses Multilayer Perceptrons (MLP) to learn higher-order combinations and nonlinear representations of features, enhancing the model's grasp on deep features of financial data and further improving the accuracy of risk prediction. Finally, DeepFM integrates the feature vectors obtained from the FM and Deep parts, and outputs the probability prediction of risks through a fully connected layer and a sigmoid function, achieving accurate assessment of financial risks. Fig. 3 shows the schematic diagram of the input vector dimension reduction process.

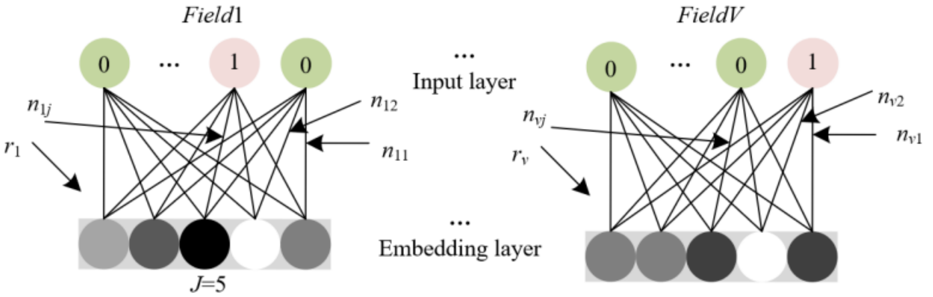

Fig. 3. Schematic diagram of the input vector dimension reduction process.

1) The FM component is specially optimized for the characteristics of financial data. This component delves into the low-dimensional feature vectors in financial data, using the factorization mechanism to identify and learn the implicit relationships and their weights between features. In this process, the FM component can adaptively adjust model parameters, effectively improving the model's performance in complex financial risk environments. The core lies in second-order feature crossing, mapping each element in the feature vector to latent factors and predicting the probability of financial risk events occurring through the interaction between these latent factors, i.e., outer product operations and summation. This approach not only significantly reduces the number of model parameters, enhancing the model's computational efficiency, but also strengthens the model's understanding and expression of complex relationships in financial feature crosses, demonstrating strong performance and practical value in the application of big data financial risk prediction and management. Assuming the bias item is represented by Q_0, the u-th component of feature vector a by a_u, the parameter of the u-th feature by q_u, the number of variables by v, and the coefficient multiplied by the u-th and k-th features by q_{uk}, the FM model is represented as follows:

$$b_{DL} = q_0 + \sum_{u=1}^{v} q_u a_u + \sum_{u=1}^{v-1} \sum_{k=u+1}^{v} q_{uk} a_u a_k \tag{16}$$

To address the issue of data sparsity in the dataset, an implicit vector for each feature is introduced, represented by $n_u=(n_{u1},n_{u2},..., n_{uj})$, and $[n_u,n_k]$ replaces Q_{uk}. At this point, the solution for Q_{uk} is transformed into the solution for n_u and n_k. The transformed formula is given as:

$$b_{DL} = q_0 + \sum_{u=1}^{v} q_u a_u + \sum_{u=1}^{v-1} \sum_{k=u+1}^{v} \langle n_u, n_k \rangle a_u a_k \tag{17}$$

This paper proposes that the weight between variables a_u and a_k can be represented by the inner product of the corresponding vectors n_u and n_k. The original complexity of FM, $P(jv^2)$, is reduced to $P(jv)$ through this transformation, assuming the inner product of vectors n_u and n_k is represented by $[n_u,n_k]$, and the d-th component of vector n_u is represented by n_{ud}, the transformation formula expression is provided as:

$$\begin{aligned} &\sum_{u=1}^{v-1} \sum_{k=1}^{v} \langle n_u, n_k \rangle a_u a_k = \\ &\quad \frac{1}{2} \sum_{u=1}^{v} \sum_{k=1}^{v} \langle n_u, n_k \rangle a_u a_k - \frac{1}{2} \sum_{u=1}^{v} \langle n_u, n_k \rangle a_{iu}^2 \\ &= \frac{1}{2} \left(\sum_{u=1}^{v} \sum_{k=1}^{v} \sum_{d=1}^{j} n_{ud} n_{kd} a_u a_k - \sum_{u=1}^{v} \sum_{d=1}^{j} n_{ud}^2 a_u^2 \right) \\ &= \frac{1}{2} \sum_{d=1}^{j} \left(\left(\sum_{u=1}^{v} n_{ud} a_u \right) \left(\sum_{k=1}^{v} n_{kd} a_k \right) - \sum_{u=1}^{v} n_{ud}^2 a_u^2 \right) \\ &= \frac{1}{2} \sum_{d=1}^{j} \left(\left(\sum_{u=1}^{v} n_{ud} a_u \right)^2 - \sum_{u=1}^{v} n_{ud}^2 a_u^2 \right) \end{aligned} \tag{18}$$

Further solved by stochastic gradient descent, the solution formula is:

$$\frac{\partial}{\partial \varphi} b(a) = \begin{array}{ll} 1, & IF\ \varphi = q_0 \\ \{ a_u, & IF\ \varphi = q_u \\ a_u \sum_{k=1}^{v} n_{kd} a_k - n_{ud} a_u^2, & IF\ \varphi = n_{ud} \end{array} \tag{19}$$

In big data financial risk prediction and management, the FM component of the DeepFM model is carefully designed for the specificity of financial data. Considering that financial data features both dense continuous variables and sparse discrete variables, the FM component optimizes data representation and storage through the concept of feature fields. After one-hot encoding, discrete data features are expanded into multiple columns forming a sparse matrix, while continuous features retain their original single-column format. To effectively address the sparsity issue caused by one-hot encoding and save storage space, the FM component introduces a transformation mechanism, converting the sparse matrix into a more compact representation, including a small dictionary of feature value indices and two small matrices: feature index matrix and feature value matrix. This design not only reduces storage requirements but also facilitates subsequent feature interaction calculations.

Fig. 4 shows the DeepFM model architecture. In the FM layer's calculation, the plus sign represents the processing of first-order features, directly associating each sparse feature part with its corresponding weight. Moreover, the cross-circle represents the calculation process of feature interactions, where green lines connect features to their dense embedding representations, calculated using the previously mentioned dictionary, feature index matrix, and feature value matrix. This process allows the model to capture complex interactions between features through low-dimensional dense vectors, particularly suited for dealing with feature-rich but sparse datasets in financial risk prediction. Through such design, the FM component of the DeepFM model not only improves computational efficiency but also enhances the model's feature expression and interaction learning capabilities when dealing with complex financial data, thus providing a more accurate and efficient prediction tool for financial risk management.

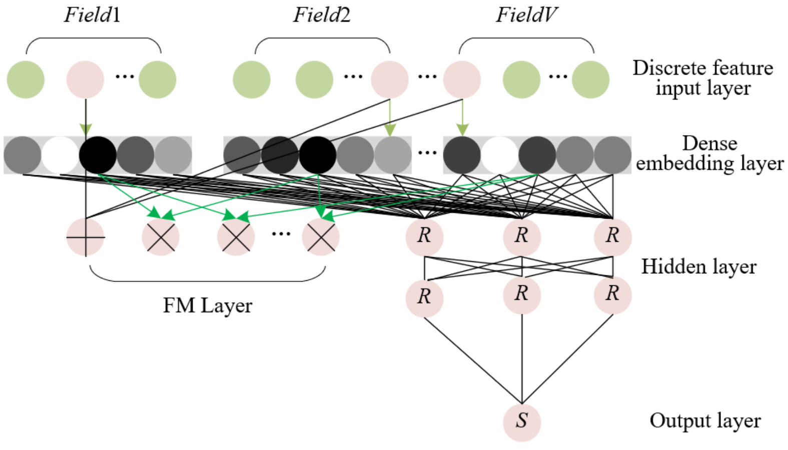

Fig. 4. DeepFM model architecture.

2) The Deep component utilizes a neural network to process and analyze high-dimensional features and complex nonlinear relationships in financial data. By introducing normalized features as inputs, the Deep component specifically targets both newly created features from feature crossing and original features, for deep semantic information extraction. This processing step is particularly important for understanding and predicting hidden patterns in financial risk. Compared to DeepFM models in other application scenarios, the model aimed at financial risk management focuses more on mining subtle and complex relationships in financial data, gradually elevating the abstract level of data features through multiple layers of fully connected layers and *ReLU* activation functions, thus capturing more refined risk signals.

Moreover, the Deep component shares the same low-dimensional dense vectors for feature embedding with the FM component, effectively enhancing the model's learning capabilities and feature expressiveness. Each feature field, regardless of its original length, is transformed into a fixed length vector *j* for uniform processing in the model. These low-dimensional vectors are then merged and input into the deep neural network for further nonlinear transformation and hierarchical feature extraction. This process ensures that the DeepFM model can maximize the use of information in financial data while maintaining computational efficiency, providing a powerful tool for financial risk prediction and management. Specifically, with $x^{(0)}= (r_1,r_2,..., r_l)$ as the output of the dense input layer and input to the hidden layers, DNN is used as the deep part of the model, assuming the number of hidden layers is represented by *G*, the output of the *m*-th hidden layer by $x^{(m)}$, and the weights and biases by $q^{(m)}$ and $y^{(m)}$, respectively, the output layer's calculation formula is given as:

$$x^{(G+1)} = \delta\left(Q^{(G)}x^{(G)} + y^{(G)}\right) \tag{20}$$

Finally, the outputs of the FM and Deep parts are integrated, with the entire process expression of the DeepFM model provided as:

$$\bar{b} = \text{sigmoid}\left(b_{FM} + b_{DE}\right) \tag{21}$$

Furthermore, effective risk management based on DeepFM model predictions can be achieved through a series of strategies. Firstly, the prediction results provide early warnings of potential risks to enterprises or financial institutions, enabling managers to adjust strategies in a timely manner, such as portfolio adjustments, optimization of loan approval processes, or implementation of risk mitigation measures, to reduce losses. By deeply analyzing risk factors and their interrelations revealed by the model, enterprises can improve their risk assessment models, developing more accurate risk rating systems. Combined with big data technology, enterprises can achieve real-time monitoring and analysis of large-scale financial data, thereby dynamically adjusting risk management strategies and improving adaptability and response speed to market changes.

IV. Experimental Results and Analysis

This paper employed a variety of financial-related datasets in its empirical analysis, including corporate financial statements, news reports, and social media comments, which are unstructured textual data. Through NLP techniques, the financial risk propensity contained within these data is extracted. Additionally, by integrating structured data such as corporate financial indicators and market data, the DeepFM model is utilized for financial risk prediction. This approach aims to verify the effectiveness of the method in enhancing predictive accuracy and efficiency. Based on the statistics of corpus labels for financial risk propensity measurement tasks shown in Table I, we can observe that model risk has the highest number of annotated corpus and sample data among all risk categories, totaling 29,691, which accounts for 65.5% of the total data. This significant amount of data not only indicates the importance of model risk in the research of financial risk prediction and management but also reflects the high demand in the market for understanding and evaluating model risk. On the other hand, liquidity risk has relatively fewer annotated corpus and sample data, totaling 290, which accounts for 0.64% of the total, suggesting that this risk category might be rare in the current dataset or relatively difficult to identify and analyze using NLP technology. This distribution indicates that there are significant differences in the attention and data availability for different risk categories when measuring financial risk propensity using NLP technology.

TABLE I. STATISTICS OF CORPUS LABELS FOR FINANCIAL RISK PROPENSITY MEASUREMENT TASKS

Risk Category	Annotated Corpus	Sample Data	Total
Market Risk	1214	714	1928
Credit Risk	2157	1025	3182
Liquidity Risk	185	105	290
Operational Risk	2241	723	2964
Compliance Risk	2895	1159	4054
Strategic Risk	1652	823	2475
Reputation Risk	465	325	790
Model Risk	21365	8326	29691
Total	32174	13200	45374

From the above data analysis, it can be concluded that research on measuring financial risk propensity using NLP is very effective in practical applications, especially when dealing with and analyzing high-frequency risk categories such as model risk. This method can process a large amount of unstructured text data, thereby revealing deep features and trends of financial risk, which is crucial for risk assessment and management. However, the research also exposes that certain risk categories, like liquidity risk, have an insufficient sample size in the current dataset, which may limit the performance and application scope of the model in these areas.

Table II presents the evaluation results of different financial risk propensity measurement methods, including several evaluation metrics such as Root Mean Square Error (RMSE), Accuracy, Precision, Recall, F1 Score, and AUC value. It can be observed from the table that the method proposed in this paper performs excellently across all metrics, especially achieving the highest in Accuracy, F1 Score, and AUC values, which are 0.9238, 0.9178, and 0.9675 respectively, while also having the lowest RMSE value at 0.2631. Compared to other popular NLP methods such as BERT, RoBERTa, GloVe, and BiLSTM-CRF, the proposed method demonstrated superior performance, particularly in handling complex financial risk prediction tasks, by accurately identifying and evaluating risks.

These results fully prove the effectiveness of the financial risk propensity measurement method based on NLP adopted in this paper. Compared to other advanced algorithms, the proposed method is more precise and reliable in extracting and analyzing unstructured textual data related to financial risks. By comparing the evaluation results of different algorithms, it can be seen that the proposed method has a clear advantage in comprehensive performance, which is particularly important in the context of financial risk prediction and management. High Accuracy and F1 Scores mean that the method can balance Precision and Recall, while a high AUC value indicates its good classification capability across different thresholds.

TABLE II. EVALUATION RESULTS OF DIFFERENT FINANCIAL RISK PROPENSITY MEASUREMENT METHODS

Method	RMSE	Accuracy	Precision	Recall	F1 Score	AUC
GRU	0.4125	0.7654	0.8546	0.6325	0.7256	0.8124
TextCNN	0.4156	0.7589	0.8236	0.6387	0.7148	0.8369
BERT	0.3256	0.8312	0.8974	0.7698	0.8156	0.9145
Self-Attention	0.3895	0.7793	0.9456	0.5841	0.7236	0.8567
Doc2Vec	0.3674	0.8215	0.8326	0.7154	0.7689	0.8576
Seq2Seq	0.3215	0.8746	0.8894	0.8756	0.8879	0.9123
LDA	0.3563	0.8312	0.8541	0.8326	0.8423	0.9178
NMF	0.3147	0.9126	0.9274	0.9124	0.9146	0.9563
BiLSTM-CRF	0.2896	0.9146	0.9236	0.9236	0.9187	0.9638
GloVe	0.2896	0.9123	0.9147	0.9157	0.9258	0.9687
RoBERTa	0.2746	0.9146	0.9133	0.9152	0.9146	0.9634
The proposed method	0.2631	0.9238	0.9186	0.9126	0.9178	0.9675

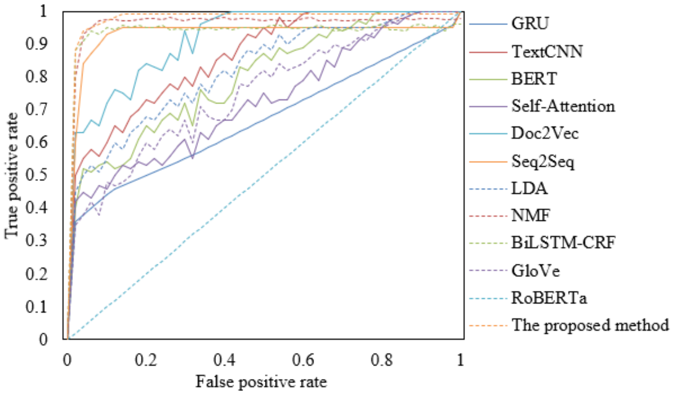

Fig. 5. Receiver Operating Characteristic (ROC) curves of different financial risk propensity measurement methods.

Analyzing the ROC curve data of different financial risk propensity measurement methods shown in Fig. 5, we can observe that the proposed method significantly outperforms other methods in performance for financial risk prediction. Especially in the area near the top right corner of the curve (close to a true positive rate and false positive rate of 1), the proposed method shows near-perfect performance, maintaining a very high true positive rate from 0 to 0.99 with almost zero false positives, ultimately achieving an optimal balance between true positive rate and false positive rate at a point close to 1. In contrast, other methods like RoBERTa, GloVe, and BiLSTM-CRF, although also showing good performance, have a noticeable gap in performance across the entire ROC curve compared to the method proposed in this paper. For example, NMF and Seq2Seq, while maintaining a higher true positive rate in most areas of the curve, still lag behind the method proposed in this paper in the capability to achieve a true positive rate above 0.99. These experimental results indicate that the financial risk propensity measurement method based on NLP not only can effectively process and analyze unstructured textual data from various channels but also has significant advantages in accuracy of risk prediction.

Comparing the indicator effects of different big data financial risk prediction methods shown in Fig. 6, we can see that the method proposed in this paper displays outstanding performance across multiple key performance indicators. Specifically, the method proposed in this paper achieves 0.93, 0.93, 0.9, 0.85, and 0.9 in Accuracy (ACC), Recall, Specificity, F1-score, and Precision respectively, which are the best or near the best performance among all compared models. Compared to other popular models like ExtraTrees, CatBoost, CNN, and Transformer, the proposed method not only shows clear advantages in prediction accuracy but also performs well in balancing Recall and Specificity, particularly achieving the highest values of 0.93 in both Recall and Accuracy, highlighting its exceptional ability to predict positive class samples. These experimental results fully validate the effectiveness of the big data financial risk prediction model based on Deep FM proposed in this paper. By deeply integrating multiple features and applying deep learning techniques, the method proposed in this paper not only improves the accuracy of predictions but also ensures the efficiency and stability of the model in processing complex and large-scale data.

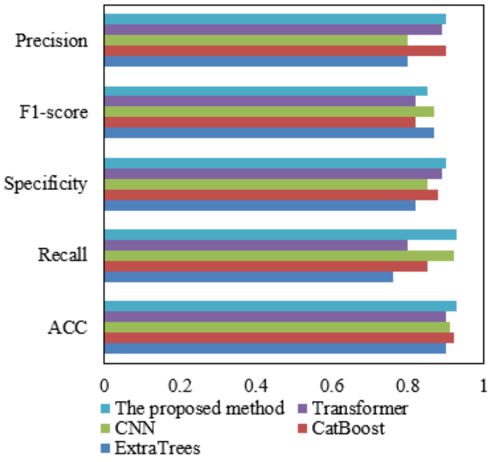

Fig. 6. Comparison of indicator effects of different big data financial risk prediction methods.

By deeply analyzing the ablation study results of big data financial risk prediction methods shown in Fig. 7, we can clearly see the advantages of the proposed method across various performance metrics. During the ablation studies, the performance of the model without optimization of the FM component and Deep component was tested separately to verify the contribution of each component to the overall model performance. When the FM component was not optimized, Accuracy (ACC), Recall, Specificity, F1-score, and Precision reached 0.902, 0.888, 0.889, 0.89, and 0.877, respectively. With the Deep component not optimized, these metrics improved to 0.912, 0.88, 0.895, 0.884, and 0.881, indicating the crucial role of the Deep component in the model. However, when both components were fully optimized, the performance of the method proposed in this paper reached its peak, with metrics of 0.925, 0.895, 0.9, 0.892, and 0.886, respectively, highlighting the importance of combining the FM and Deep components. These ablation study results clearly demonstrate the efficiency and effectiveness of the big data financial risk prediction model based on Deep FM proposed in this paper. Through the close integration of the FM and Deep components, the model not only effectively integrates various features but also improves prediction accuracy and efficiency through deep learning technology. The FM component enhances the model's understanding of feature combinations by learning interactions between features, while the Deep component captures complex nonlinear relationships through deep networks, further enhancing the model's predictive power. The success of the method proposed in this paper validates that the Deep FM-based model can provide more accurate and comprehensive risk assessments when dealing with large-scale and complex financial risk data, offering a new efficient tool for financial risk management and prediction.

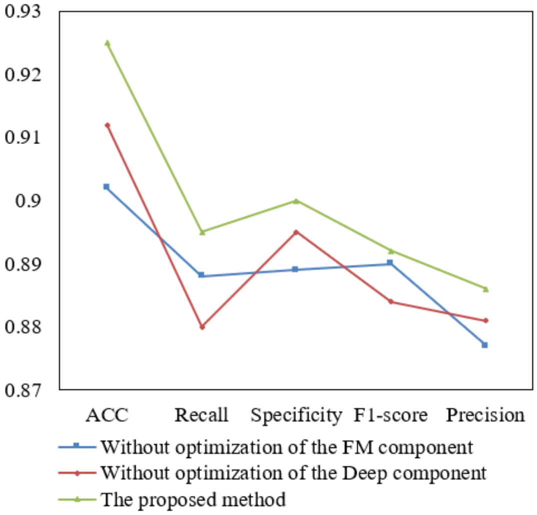

Fig. 7. Comparison of ablation study results for big data financial risk prediction methods.

The findings of this study indicated that by leveraging NLP techniques to extract key information from multi-channel unstructured text data and integrating diverse features with the deep learning capabilities of a DeepFM model, significant improvements in the accuracy and efficiency of financial risk prediction can be achieved. Specific experimental analyses demonstrated that the proposed method outperforms traditional risk measurement approaches and predictive models in terms of precision and generalization ability. Analysis of ROC curves further confirmed the notable role of the proposed method in enhancing predictive performance. Compared to existing research, our approach exhibited significant advantages in handling large-scale multi-source data and improving predictive performance, especially under conditions of high market volatility, demonstrating greater robustness and adaptability. These results validated the practical application value of our method in real-world financial risk management.

V. Conclusion

This paper applied machine learning and NLP technologies comprehensively to explore methods of predicting and managing financial risks in the context of big data. The study began with a detailed measurement of financial risk propensity using NLP technology, effectively extracting key information from unstructured text data from multiple channels, providing a richer and deeper perspective for risk assessment. Subsequently, the paper introduced an innovative financial risk prediction model based on Deep FM, which significantly improved the accuracy and efficiency of risk prediction by integrating diverse features and utilizing the powerful capabilities of deep learning.

Multiple analyses and comparisons in the experimental section demonstrated the effectiveness and advantages of the proposed method. Detailed statistics and evaluations of financial risk propensity measurement tasks showcased its deep data analysis capability. Comparative analyses of different risk measurement methods and ROC curve analyses further validated the precision and generalization ability of the proposed method. Additionally, comparisons and ablation studies of different big data financial risk prediction methods highlighted the significant role of the Deep FM model in enhancing predictive performance.

Despite achieving a series of positive results in the field of financial risk prediction and management, this paper still has certain limitations. For example, the predictive capability of the model largely depends on the quality and completeness of the data, and the processing and parsing of unstructured textual data still face challenges. Future research could explore more advanced NLP and machine learning technologies to improve the model's ability to handle complex data and enhance prediction accuracy. Additionally, the research could be expanded to more types of financial risks and explore the adaptability and stability of the model under different financial environments and conditions.

Acknowledgment

This paper was supported by the project of Hunan Provincial Natural Science Foundation (Grant No.: 2023JJ60224).

References

[1] C. Kayahan and T. Murat, "The Evolution of Financial Risk Management," J. Corp. Gov. Insur. Risk Manag., vol. 9, no. S1, pp. 155-168, 2022.

[2] P. I. Kurniawan, "Effect of Expected Return, Self Efficacy, and Perceived Risk on Investment Intention: An Empirical Study on Accounting Master Degree in Udayana University, Bali," J. Account. Financ. Audit. Stud., vol. 7, no. 1, pp. 40-55, 2021.

[3] J. Stefany and L. Agustina, "Do corporate social responsibility and political connections matter to financial performance and financial stability in the banking sector? Evidence from Indonesia," Int. J. Sustain. Dev. Plan., vol. 17, no. 8, pp. 2445-2452, 2022.

[4] D. Shen and W. Huang, "The study on commercial bank's risk management behaviour with the innovation of its scientific and technological financial product by big data analysis algorithms," International Conference on Decision Science & Management, Changsha, 2023, pp. 68-77.

[5] Q. Fan, "Risk prediction model of financial lending big data leakage based on association rules," International Conference on Decision Science & Management, Changsha, 2022, pp. 617-629.

[6] J. Xue, "Early warning of internet financial risk based on big data," 2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China, 2022, pp. 1048-1051.

[7] X. Zou, "Financial risk assessment management of state-owned enterprises based on cloud accounting in the era of big data," Appl. Math. Nonlinear Sci., vol. 9, no. 1, pp. 1-13, 2024.

[8] L. Wang, "Financial risk analysis system and supervision based on big data and blockchain technology," Secur. Priv., vol. 6, no. 2, pp. e224, 2023.

[9] T. K. Samson, "Comparative Analysis of Machine Learning Algorithms for Daily Cryptocurrency Price Prediction," Inf. Dyn. Appl., vol. 3, no. 1, pp. 64-76, 2024.

[10] S. Patalay and M. R. Bandlamudi, "Decision support system for stock portfolio selection using artificial intelligence and machine learning," Ing. Syst. Inf., vol. 26, no. 1, pp. 87-93, 2021.

[11] S. Kokate and M. S. R. Chetty, "Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model," Int. J. Saf. Secur. Eng., vol. 11, no. 5, pp. 565-572, 2021.

[12] H. Liu, "Financial risk intelligent early warning system of a municipal company based on genetic tabu algorithm and big data analysis," Int. J. Inf. Technol. Syst. Approach (IJITSA), vol. 15, no. 3, pp. 307027, 2022.

[13] J. Xiao, "Risk Control Strategy of Internet Finance Based on Financial Big Data Background," The 2021 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy, Changsha, 2022, pp. 820-824.

[14] T. Xie, "The application of financial technology in the intelligent management of credit risk under the background of big data," The International Conference on Cyber Security Intelligence and Analytics, Changsha, 2023, pp. 127-136.

[15] Z. A. Hu, "Machine learning algorithms in financial market risk prediction," Proceedings of the 2022 6th International Conference on E-Business and Internet, Singapore, 2022, pp. 301-305.

[16] Z. Wang and Y. Zhao, "Research on financial risk control model based on machine learning," 2023 2nd International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME), Madrid, Spain, 2023, pp. 313-316.

[17] M. I. Bonelli and E. S. Döngül, "Robo-advisors in the financial services industry: Recommendations for full-scale optimization, digital twin integration, and leveraging natural language processing trends," 2023 9th International Conference on Virtual Reality (ICVR), Xianyang, China, 2023, pp. 268-275.

[18] W. Y. Chen, S. H. Li, and Y. H. Wang, "Research on natural language processing in financial risk detection," Cognitive Cities: Second International Conference, IC3 2019, Kyoto, Japan, 2019, pp. 448-455.

[19] T. Magoc, K. S. Allen, C. McDonnell, J. P. Russo, J. Cummins, J. R. Vest, and C. A. Harle, "Generalizability and portability of natural language processing system to extract individual social risk factors," Int. J. Med. Inform., vol. 177, Art. no. 105115, 2023.

[20] Y. Liu, "Artificial intelligence and machine learning based financial risk network assessment model," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 158-163.