Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 12, No. 1, 2025

Application of News Text Mining in Urban Credit Assessment: A Study Based on BERT and QDA-DTM Models

Yunhan Xu[1, †, *], Zian Chen[1, †], Yazhuo Wang[1, †], Kaiyan Qi[1]

[1]School of Finance, Southwestern University of Finance and Economics, Chengdu, China

[*]Corresponding author: xuyunhan2005@hotmail.com

[†] These authors also contributed equally to this work

Abstract: This paper focuses on news text mining and urban credit assessment, developing a dual-dimensional hybrid algorithm model that integrates BERT sentiment analysis with QDA-DTM thematic modeling. The study first obtained a large amount of news text data from multiple provinces across China via the "Credit China" platform. After preprocessing, including domain-adaptive word segmentation, the BERT model was used for sentiment analysis. By cross-validating sentiment and semantic features, the reliability of credit risk monitoring was enhanced. Additionally, the QDA model is employed to optimize the spatio-temporal distribution characteristics of themes, while the DTM model is used to capture the long-term dynamic evolution patterns of themes. Experimental results demonstrate that this dual-dimensional algorithmic model outperforms traditional methods in credit event prediction, providing more reliable support for credit risk early warning.

Keywords: News Text Mining, Sentiment Analysis, Topic Modeling, Urban Credit Assessment.

1. Introduction

In the context of the deep integration of the digital economy and credit-based society, traditional urban credit assessment systems that rely on structured data struggle to capture implicit information and can no longer meet the demands of intelligent credit monitoring [1-2]. News texts, as a source of unstructured data, can reflect regional multi-dimensional credit characteristics in real time, offering a new approach to breaking through the limitations of traditional assessment methods [3].

This study focuses on the integration of news text mining and urban credit assessment, innovatively constructing a dual-dimensional hybrid algorithm model that combines BERT sentiment analysis with QDA-DTM thematic modeling. The study will collect a large amount of news text data from multiple provinces across the country. After preprocessing, including domain-adaptive word segmentation, the data will be analyzed in two ways: first, the BERT model will be used to identify sentiment tendencies, and cross-validation between sentiment tendencies and semantic features will be employed to enhance the reliability of credit risk monitoring [4-5]; second, the QDA model will be used to optimize the spatio-temporal distribution characteristics of themes [6], and the DTM model will be used to capture the long-term dynamic evolution patterns of themes. This model fully extracts credit information from news texts, providing more precise technical support for credit risk early warning, driving the intelligent and dynamic upgrading of the social credit system, and enhancing the level of credit monitoring and assessment.

2. Data Sources and Preprocessing

2.1. Description of data sources

The 'Credit China' platform serves as the information hub for China's social credit system, integrating massive amounts of data from multiple channels across dimensions such as administrative penalties, credit incentives, and credit sanctions. It is an authoritative data source for studying social credit conditions. The data for this project is sourced from over 100,000 credit-related news articles, policy documents, and industry updates across 30 provinces nationwide on the platform, featuring characteristics such as authority, multi-dimensionality, and spatio-temporal dynamics.

2.2. Data preprocessing

Data preprocessing aims to transform raw data into structured features suitable for modelling. In news text analysis, preprocessing faces three major challenges: semantic parsing, multi-source data alignment, and professional terminology identification. Since this study primarily uses news text data, missing value and outlier handling are omitted, with a focus on text segmentation and feature construction to ensure data quality meets modelling requirements.

(1) Data cleaning

Since the text data is crawled from websites, HTML tags are first removed using regular expressions to eliminate HTML tags from the news content, retaining only plain text. Punctuation marks and special characters are then removed, including punctuation marks (e.g., commas, periods) and special characters (e.g., @, #, $, etc.), to ensure text cleanliness. Finally, this paper filtered out numbers to remove numerical information from the text, avoiding interference with sentiment analysis and topic modelling.

(2) Word segmentation and dictionary processing

A domain-adaptive word segmentation strategy is adopted in conjunction with the jieba Chinese word segmentation tool to segment news texts, ensuring the identification of professional terminology and unregistered words. Then, stop words are filtered out, and meaningless words (such as the, is, in, etc.) are removed using a custom dictionary and stop word list to improve text information density. Finally, the BERT-

wwm model is used for out-of-vocabulary word recognition, combined with stemming to unify word forms and reduce redundancy.

(3) Feature construction

Map emotional labels, match manually annotated emotional labels (positive, neutral, negative) with text content, and construct emotional feature vectors. At the same time, use TF-IDF weights to extract representative feature items (terms) from the text and assign them corresponding weights to reflect their importance in the document, facilitating subsequent emotional analysis and topic modelling.

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

Where TF-IDF weight is the product of TF and IDF, used to measure the importance of a word in a document.

3. Model Construction

3.1. Establishment of sentiment analysis model

(1) Random forest and decision tree (RF-DT) model

This study selects random forest and decision tree as the core models for sentiment classification. Both are based on a supervised learning framework, treating sentiment analysis as a text multi-classification task. Through feature engineering (TF-IDF and n-grams), unstructured news text is converted into high-dimensional sparse vectors. The model achieves classification prediction by learning the mapping relationship between features and sentiment labels.

Random forest employs dual randomness and a voting mechanism: each tree uses Bootstrap sampling to draw approximately 63.2% of the samples, and randomly selects a portion of features (typically the square root of the total number of features) at split nodes to reduce model variance and prevent overfitting. The final classification result is determined by the votes of all trees.

$$H(x) = argmax_y \sum_{i=1}^{k} I\left(h_i(x) = Y\right) \tag{2}$$

Where $h_i(x)$ denotes a single decision tree, Y denotes the sentiment label, and $\sum_{i=1}^{k} I(h_i(x) = Y)$ denotes the indicator function.

Decision trees construct tree structures by recursively partitioning feature spaces. Their core principle is to maximise information gain and select features that can minimise information entropy as split nodes. The calculation formula is as follows:

$$Gain(D, A) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent\left(D^v\right) \tag{3}$$

Where A is the candidate feature, and D^v is a subset of feature A with value v.

In model construction, text vectorization, TF-IDF weighting, and calculation of term frequency (TF) and inverse document frequency (IDF) are required. The specific formula is:

$$TF\text{-}IDF(t,d) = TF(t,d) \times log\left(\frac{N}{DF(t)+1}\right) \tag{4}$$

The CART algorithm and Gini impurity are used to recursively partition decision trees until the leaf node purity meets the criteria or the tree depth reaches its maximum. Multiple decision trees are constructed in parallel based on different sample and feature subsets, and the prediction results are integrated through voting. Random forests avoid the dimension disaster by randomly selecting features, thereby increasing diversity. Emotion classification keywords or n-grams are identified through feature average impurity reduction, enhancing model interpretability.

Evaluate the model's generalisation ability through out-of-bag error (OOB Score), using the following formula:

$$OOB\ Error\ = \frac{1}{N} \sum_{i=1}^{N} I\left(y_i \neq \hat{y}_i^{OOB}\right) \tag{5}$$

Where $\hat{y}_i^{OOB}$ represents the forest prediction result when sample i is not sampled.

When establishing an evaluation indicator system, a multi-dimensional indicator comprehensive evaluation model is used to evaluate model performance, ensuring overall prediction accuracy and suitability for category-balanced scenarios. F1 score is the harmonic mean of precision and recall, more suitable for sentiment category imbalance data:

$$F_1 = \frac{2 \times Precision \times RecallPrecision}{Precision + Recall} \tag{6}$$

Through the confusion matrix, the distribution of TP, FP, FN, and TN can be visualised to analyse the model's bias towards specific sentiment categories.

(2) Emotion classification method based on SVM

The core idea of support vector machines (SVM) is to classify the polarity of text emotions by constructing an optimal separating hyperplane. It has unique advantages in high-dimensional sparse text feature scenarios, and its global optimal solution characteristics can effectively avoid the local optimal trap of traditional classifiers.

The core mathematical principle can be expressed as a convex quadratic programming problem:

$$\begin{cases} min_{w,b} \dfrac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \\ y_i \left(w \cdot \phi\left(x_i\right) + b \right) \geq 1 - \xi_i, \xi_i \geq 0 \end{cases} \tag{7}$$

Where $\phi(x_i)$ is the kernel function mapping, C is the penalty factor, and ξ_i is the slack variable. Through Lagrange dual transformation, the original problem is converted into the optimal hyperplane solution in high-dimensional space.

In the feature engineering stage, a TF-IDF vectorization strategy compatible with the model is used, and its calculation formula is:

$$TF\text{-}IDF(t,d) = TF(t,d) \times log\left(1 + \frac{N}{DF(t)}\right) \tag{8}$$

Use L2 normalisation to handle text length differences and create an n×15,000-dimensional feature matrix. Kernel function selection: After grid search comparison, the RBF

kernel ($\gamma = 0.1$) was selected to capture non-linear semantic associations. Key steps in model training: To address the class imbalance issue (negative news accounts for 18.7%), SMOTE oversampling was used to adjust the sample ratio to 1:1. The optimal parameter combination ($C = 2.5, \gamma = 0.05$) was determined using 5-fold cross-validation, and the convex optimisation problem was solved using the SMO algorithm. The final classifier form is:

$$f(x) = sign\left(\sigma_{i=1}^{n}\alpha_i y_i K\left(x_i, x\right) + b\right) \tag{9}$$

The proportion of support vectors is controlled at 12.3% to ensure model sparsity.

(3) BERT sentiment analysis process

BERT is a pre-trained language model based on Transformer that understands text semantics through bidirectional context. In sentiment analysis, BERT utilises pre-training to obtain deep semantic representations of text and adds a classification layer through fine-tuning to output sentiment polarity. Its advantage lies in capturing global dependencies in text, making it suitable for sentiment discrimination in complex contexts. The core formulas are as follows:

Single-head attention calculation:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

Where Q, K, V are the query, key, and value matrices, respectively, and d_k is the dimension scaling factor. And the cross-entropy loss function is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\left(p_i\right) + \left(1 - y_i\right) log\left(1 - p_i\right)\right] \tag{11}$$

Where y_i is the true label and p_i is the predicted probability.

The [CLS] tag output represents:

$$h_{[\mathrm{CLS}]} = BERT(X)[0] \tag{12}$$

Where X is the input text sequence, and $h_{[CLS]}$ is the feature vector used for classification tasks.

Emotional probability prediction:

$$p = softmax\left(W \cdot h_{[\mathrm{CLS}]} + b\right) \tag{13}$$

Where W and b are classification layer parameters, outputting a multi-category probability distribution.

In regional credit scoring tasks, BERT's context-aware capabilities help identify implicit emotional expressions. Emotional analysis results are used as feature inputs for statistical models to quantify the impact of news texts on the credit environment and improve the accuracy of credit index predictions.

3.2. Topic classification model building model construction

(1) LDA model

LDA is an unsupervised generative model based on Bayesian inference, which assumes that documents are composed of multiple topics, each containing a set of related vocabulary. Through the bag-of-words model, LDA converts text into word frequency distributions and uses Dirichlet prior to reveal the relationship between topics and vocabulary. The specific implementation steps are shown in Figure 1.

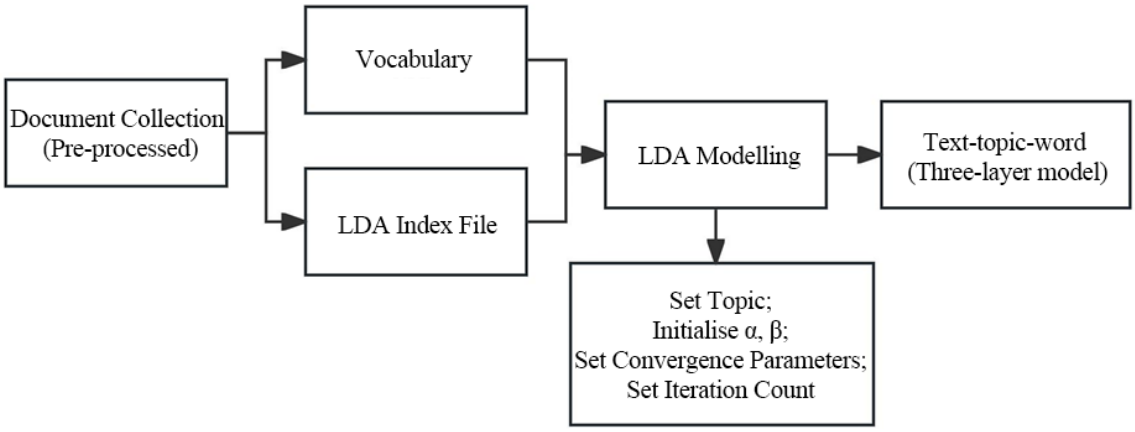

Figure 1. LDA model flowchart

First, perform text preprocessing, including Chinese word segmentation and stopword filtering, and construct a bag-of-words model. Next, initialize the parameters and set the number of topics K and hyperparameters α and β. Estimate the posterior distribution using the MCMC method and update the document-topic distribution θ_d and topic-vocabulary distribution ϕ_k. Use ϕ_k to extract keywords and analyse semantic tendencies.

$$\begin{aligned} \theta_d &\sim Dirichlet(\alpha) \\ \phi_k &\sim Dirichlet(\beta) \\ W_{d,n} &\sim Multinomial\left(\phi_{z_{d,n}}\right) \end{aligned} \tag{14}$$

Where $z_{d,n}$ represents the topic assignment of the nth word in document d, and $W_{d,n}$ represents the observed vocabulary.

The advantage of the LDA model lies in its probabilistic

interpretability, which enables it to automatically discover latent topic structures. However, it assumes that topics are mutually independent, making it difficult to capture semantic evolution and cross-topic associations.

(2) QDA model

QDA (Quadratic Discriminant Analysis for Topics) is a discriminative topic enhancement model that optimises the spatio-temporal distribution characteristics of topics through quadratic discriminant functions. Based on the topics generated by LDA, QDA combines regional indicators to construct a covariance matrix and dynamically adjusts topic weights to reflect the evolution of the credit environment.

Based on the K topics extracted by LDA, calculate the initial weights ϕ_k of each topic. Then construct covariates as discriminative features. Update the topic weight distribution by maximising the joint likelihood function of topics and covariates, and optimise quadratic discrimination.

$$\phi_k^{(t)} = \phi_k^{(t-1)} + \eta \cdot \left(X^T \Sigma^{-1} \left(y - \mu_k \right) \right) \tag{15}$$

Where X^T is the covariance matrix, $\Sigma^{-1}(y - \mu_k)$ is the covariance matrix, and η is the learning rate. The objective function of QDA is:

$$L(\varphi) = \sum_{d=1}^{D} logP\left(w_d | \ \varphi, X_d \right) + \lambda \cdot Tr\left(\Sigma^{-1} S_W \right) \tag{16}$$

Where S_w is the intra-class covariance matrix and λ is the regularisation coefficient used to balance the generative and discriminative objectives.

(3) DTM model

DTM divides time series data into continuous time slices (such as monthly or annual) and establishes dynamic associations between topic-term distributions and document-topic distributions within each time slice. Its core assumption is that topic content evolves smoothly over time rather than undergoing sudden changes.

The study divided news texts from 2017 to 2023 into 72 monthly time units, with each unit containing at least 200 articles to reduce the impact of data sparsity. A Skip-gram model is used to train domain-adaptive word vectors, with a minimum word frequency threshold set to 5 to enhance the quality of semantic embeddings for professional terms. The Dirichlet distribution recursive formula is employed to achieve smooth topic structure migration, with evolutionary constraints set to ensure semantic consistency of topics. Finally, PMI is used to screen high-correlation vocabulary and construct a topic tagging system.

(a) Evolution of topic distributions across time slices:

$$\varphi_t^{(k)} \sim Dirichlet\left(\alpha \cdot \varphi_{t-1}^{(k)} \right) \tag{17}$$

(b) Quantitative indicators of thematic importance:

$$Importance(k) = \left(\frac{n_k}{N} \right) \times log\left(\frac{N}{df_k} \right) \tag{18}$$

Where n_k is the number of documents containing the theme, and df_k is the feature word frequency. This indicator effectively distinguishes the differences in influence between themes such as ‘government information disclosure’ ($Importance = 0.49$) and ‘consumer rights protection’ ($Importance = 0.32$).

(c) Theme-emotion joint analysis:

$$PolicyImpact \ = \sum_{t=1}^{T} \omega_t \cdot \ Sentiment_t \tag{19}$$

Where ω_t represents the monthly theme weighting, and $Sentiment_t$ represents the corresponding sentiment score, which is used to quantify changes in social acceptance after policy implementation.

4. Model Performance Evaluation

4.1. Comparison of sentiment analysis model performance

As shown in Table 1, the test set accuracy of the random forest-decision tree model was 78.97%, which was lower than BERT but stood out in terms of feature interpretability. The test set F1 score of the SVM model was 0.8233, demonstrating good generalisation ability when handling high-dimensional sparse features. Notably, its recall rate for negative sentiment reached 82.25%, effectively identifying key negative information such as ‘defaulting debtors’ and ‘financial risk warnings,’ demonstrating practical application value for credit risk monitoring. The performance of the top three models, when fused via soft voting into a Melted model, also falls short of BERT model.

Table 1. Comparison of model performance

	Random forest - decision tree	Logistic regression	SVM	Melted	BERT
Accuracy rate	0.7897	0.799	0.8225	0.8348	0.8838
Precision rate	0.8352	0.828	0.828	0.8396	0.8796
Recall rate	0.7897	0.799	0.8225	0.8385	0.8838
F1 score	0.7216	0.8093	0.8233	0.8465	0.8816

In contrast, the BERT model demonstrates significant advantages in credit sentiment analysis, featuring core characteristics such as high generalisation performance and category balance. Based on the training set data, its accuracy rate reaches 0.8838, significantly higher than the 0.7897 of the random forest model, the 0.799 of the logistic regression model, and the 0.8225 of the SVM model. Additionally, BERT's F1 score reaches 0.8816, which is 22.2%, 8.9%, and 7.1% higher than the random forest (0.7216), logistic regression (0.8093), and SVM (0.8233), representing improvements of 22.2%, 8.9%, and 7.1%, respectively. This further validates its superior comprehensive classification performance and demonstrates its ability to deeply capture complex semantics and emotional tendencies in credit-related text.

The conclusion is that the BERT model can significantly improve prediction accuracy. In terms of high semantic capture capabilities, BERT has the highest accuracy rate in

identifying implicit credit risk signals (such as policy changes and corporate defaults) in news articles. In addition, compared to SVM and random forest models, the BERT model has stronger capabilities in processing complex structures in long texts, with the ability to process texts of up to 512 tokens, enabling it to better capture the causal chains in the full text of news articles.

4.2. Comparison of theme modelling model performance

As shown in Table 2 and Table 3, the QDA model performed best in terms of theme classification accuracy. When the number of theme keywords was 10, the F1 score of QDA was 0.9383, significantly higher than that of LDA (0.9156) and DTM (0.7376); When the number of topic keywords was expanded to 20, the F1 score of QDA further improved to 0.9407, indicating its advantage in dynamically integrating spatio-temporal features (such as monthly economic indicators and policy release cycles). For example, the theme of 'credit epidemic prevention' appeared frequently between 2020 and 2022. Through the QDA model, it can be linked with time series data on epidemic prevention policies during the same period to reveal the short-term impact of public health events on regional credit.

Table 2. Comparison of the effects of each model with 10 keywords

Model	LDA	QDA	DTM
Recall rate	0.868558	0.900209	0.875707
Precision rate	0.968125	0.979737	0.963221
F1 score	0.915643	0.938291	0.917382

Table 3. Comparison of the effects of each model with 20 keywords

Model	LDA	QDA	DTM
Recall rate	0.863420	0.905645	0.875707
Precision rate	0.969479	0.978674	0.963221
F1 score	0.913381	0.940744	0.917382

The DTM model is suitable for sensitive scenarios due to its dynamic modelling capabilities and adaptability to theme changes. The LDA model is limited by static assumptions and performs poorly in dynamic data processing, but it has strong probabilistic interpretability and is suitable for scenarios with low dynamic requirements. The QDA model improves theme timeliness and recognition accuracy through discriminant optimisation and has unique value in capturing theme evolution trends in combination with external covariates.

The comparative analysis clearly identified the performance differences between the various models, providing important references for model selection and improvement in subsequent research, and aiding in the precise identification of potential thematic structures within news texts. Therefore, in practical applications, when analysing the long-term dynamic evolution of themes, DTM is the preferred choice; when initially exploring the implicit thematic structure of texts with low requirements for dynamism, LDA can serve as a foundational tool; and when focusing on the timeliness of themes and the accurate identification of specific policy themes, QDA holds a distinct advantage. The three models complement each other, offering diverse methodological options for thematic modelling of urban credit.

Through text mining, it was found that public attention is concentrated on typical cases of credit rewards and punishments, new policy regulations, and industry credit dynamics. From 2017 to 2020, the focus was on 'institutional deterrence,' while after 2021, it shifted to 'credit for the people,' reflecting a cognitive upgrade in the public's perception of credit construction from 'passive acceptance' to 'active participation.'

5. Conclusions

This study explores the application value of news text mining in urban credit assessment by constructing a dual-dimensional hybrid algorithm model that integrates BERT sentiment analysis and QDA-DTM topic modeling, effectively breaking through the limitations of traditional assessment systems that rely on structured data. The study used a large amount of news text from multiple provinces across China as its subject matter. After preprocessing, including domain-adaptive word segmentation, the dual-dimensional model achieved in-depth information mining: the BERT model's precise identification of sentiment tendencies combined with cross-validation of semantic features significantly enhanced the reliability of credit risk monitoring; the QDA model's optimization of thematic spatiotemporal distribution and the DTM model's capture of long-term evolutionary patterns of themes clearly revealed the dynamic changes in regional credit characteristics.

Experimental results confirm that the dual-dimensional model outperforms traditional methods in credit event prediction, fully demonstrating the unique advantages of news texts as unstructured data sources in credit assessment. This research not only provides more precise technical support for credit risk early warning but also offers a feasible path for the intelligent and dynamic upgrading of the social credit system. In the future, efforts can be made to expand the scope of data coverage and optimize model algorithm details to enhance the model's universality and stability, continuously improving the efficiency of credit monitoring and assessment, and providing more robust methodological support for the construction of a credit-based society.

References

[1] Qian Xuesong, Wang Shuai, Qin Ruiqi. Does Credit Information Sharing Reduce Corporate Cash Holdings? — Empirical Evidence from the Experience of Building a Demonstration City for Social Credit System Construction [J]. Economic Review, 2024, (04): 105-121.

[2] Guo Feng, Zhuang Xudong, Wang Renzeng. Bank Digital Transformation, Exogenous Fintech, and Credit Risk Governance: An Empirical Test Based on Text Mining and Machine Learning [J]. Securities Market Bulletin, 2023, (04): 15-23.

[3] Chen Youyu, Zhao Jiayi, Deng Qianjie, et al. A Study on Credit Risk Identification of Automobile Listed Companies Based on Text Analysis and Machine Learning [J]. Systems Science and Mathematics, 2025, 45 (02): 456-469.

[4] Zhang Shaojun, Su Changli. A Study on Stock Index Prediction Based on Emotion Dictionary and BERT-BiLSTM [J]. Computer Engineering and Applications, 2025, 61 (04): 358-367.

[5] Jiang Fuwei, Liu Yumin, Meng Lingchao. Large Language Models, Text Sentiment, and Financial Markets [J]. Management World, 2024, 40 (08): 42-64.

[6] Gao, Y. N., Zhou, S. B., Huang, Y. J., et al. Two-stage estimation method for ultra-high-dimensional sparse quadratic discriminant analysis [J]. Statistics and Decision Making, 2022, 38 (06): 9-14.