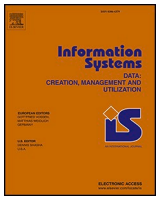
Information Systems 128 (2025) 102484

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is



Emerging industry classification based on BERT model

Baocheng Yang [a], Bing Zhang [a], Kevin Cutsforth [b,*], Shanfu Yu [a], Xiaowen Yu [c]

[a] Huanghe Science and Technology University, Zhengzhou, PR China
[b] Royal Agriculture University, UK
[c] Henan Finance University, Zhengzhou, PR China

ARTICLE INFO

Keywords:
Industry classification
Machine learning
BERT

ABSTRACT

Accurate industry classification is central to economic analysis and policy making. Current classification systems, while foundational, exhibit limitations in the face of the exponential growth of big data. These limitations include subjectivity, leading to inconsistencies and misclassifications. To overcome these shortcomings, this paper focuses on utilizing the BERT model for classifying emerging industries through the identification of salient attributes within business descriptions. The proposed method identifies clusters of firms within distinct industries, thereby transcending the restrictions inherent in existing classification systems. The model exhibits an impressive degree of precision in categorizing business descriptions, achieving accuracy rates spanning from 84.11% to 99.66% across all 16 industry classifications. This research enriches the field of industry classification literature through a practical examination of the efficacy of machine learning techniques. Our experiments achieved strong performance, highlighting the effectiveness of the BERT model in accurately classifying and identifying emerging industries, providing valuable insights for industry analysts and policymakers.

1. Introduction

Precise industry categorization serves as a cornerstone in the analysis of industrial structure and dynamics. It is also central to broader disciplines in economics and education departments, where the study of industries is essential for drawing valid statistical inferences from empirical samples. Comprehending the economic landscape and its evolution within a region or nation is paramount. Industry classification serves as a tool for governments and policymakers to tailor policies effectively across various sectors. Enterprises leverage industry classification to perform market analyses and pinpoint avenues for expansion.

According to the National Standard of Classification of National Economic Industries, China is divided into 20 categories, 97 major categories, 473 medium categories and 1380 subcategories. In addition, the National Bureau of Statistics has successively formulated, revised, and promulgated 16 classification standards for derivative industries Table 1

While existing industry classifications offer convenience for economic planning and research, they have limitations. The methods are outdated and inefficient. The current system makes it difficult to determine if an entity belongs to one of the 16 derivative industries. Counting and subdividing these industries is even more challenging, requiring significant labor and material resources. Additionally, they do not possess ongoing metrics for gauging similarity both within individual industries and across different sectors. Timely updates to reflect the evolving business and industry landscape are another concern. Furthermore, the application of new technologies is driving significant advancements in social productivity, leading to profound changes in industrial and economic structures. While many new industries have emerged in practice, current industry classifications inadequately capture these future trends.

Motivated by the limitations of existing industry classification methods, this work proposes solutions using the Bidirectional Encoder Representations from Transformers (BERT) model. We leverage the power of a pre-trained, domain-specific BERT model to present a simple and highly efficient approach for industry classification in a Chinese context. Our key contributions are:

1. This study introduces a novel machine learning approach to industry classification.
2. We investigate the effectiveness and accuracy of domain-specific BERT embeddings for classifying large volumes of text data.

* Corresponding author.
E-mail address: kevin.cutsforth@rau.ac.uk (K. Cutsforth).

https://doi.org/10.1016/j.is.2024.102484
Received 14 October 2024; Accepted 15 October 2024
Available online 16 October 2024

Table 1
15 industry classification standards developed by the National Bureau of Statistics.

Serial number	Name	Date of enactment	Date of revision
1	Classification of Culture and Related Industries	2012	2018
2	Classification of High-tech Industries (Service industry)	2013	2018
3	Classification of High-tech Industries (Manufacturing)	2013	2017
4	Classification of Health Service Industry (Trial)	2014	Statistical Classification of Health Industry, 2019
5	National Statistical Classification of Science and Technology Service Industry (2015)	2015	2018
6	National Statistical Classification of Tourism and Related Industries	2018	
7	Classification of Strategic Emerging Industries	2018	
8	Statistical Classification of Producer Services	2019	
9	Statistical Classification of Life Services	2019	
10	Statistical Classification of Sports Industry	2019	
11	Statistical Classification of the Elderly Care industry	2020	
12	Classification of Education, Training and Related Industry Statistics	2020	
13	Statistical Classification of Agriculture and Related Industries	2020	
14	Statistical Classification of Digital Economy and Its Core Industries	2021	
15	Statistical Classification of Energy Conservation, Environmental Protection, and Clean Industry	2021	

3. We perform a series of in-depth experiments to assess the effectiveness of a pre-trained BERT model, demonstrating its adaptability to different challenges through fine-tuning for specific industries.

This paper is structured to elucidate the key aspects of our research. Section 2 delves into a comprehensive review of relevant literature, establishing the foundation for our work. Section 3 formally defines the specific task under investigation, ensuring clarity in our research focus and meticulously details our proposed model, outlining its components and functionalities. Section 4 presents the experimental design, execution, and a rigorous analysis of the obtained results. Finally, Section 5 concludes the paper by summarizing our findings and contributions to the field.

2. Related studies

Earlier research in this field focuses on data-driven industry classification. Hoberg and Phillips [1] constructed a network industry classification system using textual analysis of companies' business descriptions from 10-K filings. Their approach demonstrated superior performance over SIC and NAICS in clustering firms with comparable products and services, it has become a widely used resource for many researchers.

Lee et al. [2] employed internet co-searches sourced from the U.S. Securities and Exchange Commission (SEC) website to identify clusters of comparable firms, positing that users of the SEC data-retrieval website jointly seek similar firms to facilitate their investment endeavors. While this approach integrates investor sentiment, its utility is constrained by the accessibility of search data. Dalziel et al. [3] advocated for the development of an industry classification framework based on inter-firm transaction data. Nonetheless, the endeavor proved unsuccessful in generating any industry groupings as a result of insufficient data availability.

2.1. Methods employing machine learning for industry categorization

Recent advancements in machine learning (ML) and natural language processing (NLP) offer promising solutions to address the limitations of traditional company classification standards. These advancements can potentially reduce costs, complexity, and manual labor. Oyarzun [4] conducted a study employing text mining and machine learning techniques aimed at furnishing Statistics Canada with a tool for coding missing industrial classifications, thereby enhancing the overall quality of classifications within the Business Register.

Following the text mining of North American Industrial Classification System (NAICS) codes, the study investigated whether a multivariate Bernoulli naive Bayes classifier could accurately forecast North American Industry Classification System (NAICS) codes by analysing features extracted from textual business variables like name, description, and main activity. Dolphin et al. [5] propose a multimodal neural network model for industry classification in finance. They utilize text embeddings generated from historical pricing data and financial news to train a support vector classifier. This study highlights the usefulness of the embeddings through case studies and their application to industry classification. This method offers an objective and data-driven approach compared to traditional, subjective classification schemes.

Building on prior work, Kim et al. [6] also leverage NLP for industry classification. They extract distinctive features from business descriptions within financial reports. Rizinski et al. [7] propose a method for company classification that leverages natural language processing (NLP) and zero-shot learning. They utilized pre-trained transformer models to extract features from company descriptions and applied zero-shot learning to categorize companies without needing specific training data for each category. By leveraging NLP, Rizinski et al. [7] achieve a more streamlined and potentially more accurate company classification process compared to GICS, which often requires manual effort.

In more recent studies, The Data City (2024) presents a real-time industry classification system (RTIC) as an alternative to traditional Standard Industrial Classification (SIC) codes. The RTIC identifies and categorizes companies based on their actual activities and data usage, enabling more granular and dynamic industry classification for various applications. Bonne et al. [8] developed a data-driven peer grouping system that employs artificial intelligence (AI) tools to capture market perception and cluster companies at multiple levels of granularity. Tagarev et al. [9] applied several text classification techniques, based on both deep learning and classical vector space models, to address the task of categorizing companies within industry classification schemes.

These studies showcase the diverse applications of machine learning for industry classification across various sectors. They highlight the ongoing research efforts to improve accuracy, efficiency, and model selection to meet specific industry needs. (Emphasis on industry-specific needs)

2.2. BERT model for industry classification

The Bidirectional Encoder Representations from Transformers (BERT) model has become a popular tool for automatic industry classification due to its ability to understand the nuances of language and context. The field of using BERT for industry classification is rapidly evolving, with new research published regularly. Here, we highlight some recent studies showcasing significant advancements:

Slavov et al. [10] explore different approaches for classifying companies using industry classification schemes. To identify the most effective model, Slavov et al. (2019) compare the performance of several pre-trained English language models (BERT, XLNet, GloVe, and ULMFit) on a semi-controlled target classification system built from DBpedia open data, two simple perceptron models serve as the baseline. Their results demonstrate that BERT and XLNet outperform other methods for multi-label classification of DBpedia company abstracts, even with unbalanced classes. This work provides an early example of using BERT for industry classification and highlights its potential for this task.

Research by Gao et al. [11] explores industry classification using BERT and various word embedding schemes (e.g., word2vec, doc2vec) and clustering algorithms (e.g., greedy cosine similarity, k-means, Gaussian mixture model). Their findings can inform the selection of appropriate methods for this task. Wang et al. [12] constructed a dataset of 17,604 annual business reports and corresponding industry labels for companies listed on China's National Equities Exchange and Quotations (NEEQ). They explored the combination of BERT embeddings and graph neural networks for improved classification accuracy over traditional methods. Their proposed framework utilizes BERT to create graph representations of industry data and train a graph neural network for classification.

Similarly, Xu and Ji [13] utilize short text classification algorithms and convolutional neural networks (CNNs), considering the nuances of various industries. They propose an industry classification algorithm based on an enhanced BERT model, which integrates CNNs with a three-channel approach to analyse a listed company's main business description at the word, phrase, and concept levels, this analysis helps determine the company's industry affiliation. In the same vein, Ito et al. [14] propose learning vector representations of companies based on their annual reports. They introduce a multi-task learning strategy that involves fine-tuning the BERT language model on two tasks: (i) classifying companies based on existing industry labels and (ii) predicting their stock market performance. Experiments conducted on a newly constructed dataset of US and Japanese companies (in English and Japanese, respectively) demonstrate the usefulness of this strategy.

A contemporary study by Jagrić and Herman (2024) demonstrates the effectiveness of BERT for classifying business descriptions into specific industries. This research provides an in-depth analysis of leveraging BERT for multi-class text classification with a focus on fine-grained industry classification, encompassing thirteen distinct categories, this study achieved an impressive 88% accuracy in classifying businesses using BERT and website data. These results highlight BERT's ability to capture nuanced industry-specific information.

Overall, the literature review shows that BERT is a powerful tool for industry classification. BERT-based models have been shown to outperform traditional machine learning techniques and other feature extraction methods. These studies showcase the growing interest and promising applications of BERT for industry classification in various domains. Recent research paints a positive picture for using BERT in industry classification. Its ability to handle complex text data and adapt to various domains makes it a valuable tool for businesses and organizations needing to categorize companies or information.

3. Data and BERT model construction

3.1. BERT model

Bidirectional Encoder Representations from Transformers (BERT) is a pre-training technique in Natural Language Processing (NLP) introduced by Google in 2018. Since its introduction, A powerful language model, BERT has set new benchmarks in various natural language understanding tasks. Its secret lies in pre-training on massive amounts of unlabelled text, followed by fine-tuning with labelled data specific to each NLP task [15].

Recent studies (e.g., [11,13,16]) have shown BERT to be a powerful tool for industry classification tasks. Its ability to understand the context and relationships between words makes it well-suited for analyzing textual descriptions of companies and accurately identifying their industry affiliation. The BERT industry classification framework involves four steps: pre-training, text processing, fine-tuning for industry classification, and classification.

During pre-training, the core of the process is a pre-trained BERT model, such as Bert-base-uncased or Bert-base-multilingual-uncased. These models have already been trained on massive amounts of text data, allowing them to understand the relationships between words and capture the meaning of sentences. The second step involves preparing the text data for classification. The pre-processing stage often incorporates text normalization techniques, such as the removal of extraneous characters (punctuation, special symbols) and tokenization into individual words or even subword units (morphemes or character n-grams).

In the third step, fine-tuning for industry classification, the pre-trained BERT model is then fine-tuned on a dataset of text labeled specifically with industry categories. This dataset could include company descriptions, news articles, financial reports, industry reports, press releases, or other relevant documents. During fine-tuning, the model is fine-tuned to identify the specific features and patterns that distinguish different industries. Finally, in the classification step, once fine-tuned, the model can be used to classify new text data. When presented with a piece of text, the model analyzes it and assigns a confidence score to each industry category. The text is then classified as belonging to the category with the highest confidence score.

3.2. Data processing and model construction

The initial dataset contained 13,887 data points related to the health industry, drawn from a larger database of 80,000 companies. After testing and refinement, the final dataset used for training comprised 231,335 data points from a 200,000-company database. The model was trained on data including company name, company profile, and company business scope. To address the lengthy nature of company profiles and business scopes, the text was pre-processed using Chinese word segmentation technology. This process reduced the data to a set of 256 high-frequency words, which were then used for training, testing, and

Table 2
Examples of the dataset.

Business description	Industry class
Communications, electronic components, chips, and terminal applications. This includes upstream base station radio frequency and baseband chip development, midstream network construction and planning, design and maintenance, as well as downstream product applications such as cloud computing, vehicle networking, Internet of Things (IoT), VR/AR scenarios. The ecosystem involves basic network equipment vendors, wireless network providers, mobile virtual network operators (MVNOs), network planning/maintenance firms, application service providers, end users.	5G
Information technology hardware and software products along with information technology service providers. Infrastructure security and endpoint security among others; while the services include security scheme integration & operation/maintenance.	Cybersecurity
Material basis & technical support for energy/resource conservation & circular economy development while protecting ecological environment. Its six areas comprise: energy-saving technology/equipment/products/services; advanced environmental protection technology/equipment/products/services.	Energy Saving and Environmental Protection

validation purposes. Table 2 displays an illustration of the transformed dataset.

A random sample of 6056 data points was selected from a large dataset within a specific industry. After cleaning and segmenting the company profile and business scope data using Chinese word segmentation technology, the data was further divided: 5056 data points were chosen for the training dataset, 500 for the test dataset, and 500 for the validation dataset.

3.2.1. Model construction

The BERT model was chosen for industry classification. BERT is a pre-trained model, meaning it has already been trained on a massive dataset of text. This pre-training allows the model to learn general features about language that can be applied to new tasks. In fine-tuning, a pre-trained model like BERT is adapted to a specific task. Here's how it works: Imagine a training dataset A. We first use A to train the network, essentially teaching it to recognize patterns relevant to task A. These learned patterns are then saved for later use. When a new task B arises, we can leverage the pre-trained model's knowledge by initializing the network with the parameters learned from A. While some high-level parameters (those controlling broader aspects of language processing) might be randomly initialized, the core knowledge from A provides a strong foundation. Finally, the network is fine-tuned with the training data specific to task B. During this fine-tuning stage, the loaded parameters from A are either frozen (remaining unchanged) or further adjusted to better suit the specific needs of task B.

3.2.2. Benefits of fine-tuning

Fine-tuning is particularly beneficial when dealing with smaller datasets like task B. Since the model already possesses a wealth of knowledge from the pre-training stage, it can leverage this foundation to learn from a smaller dataset more effectively compared to training a model from scratch Fig. 1

This figure illustrates the two main stages involved in training a BERT model: pre-training and fine-tuning.

- Pre-training: The pre-training stage is a multi-task learning process that utilizes two tasks:
- Masked Language Modeling (MLM) constitutes a pre-training objective where a proportion of words within the input sequence are masked, and the model is trained to predict the masked elements based solely on the contextual information gleaned from the surrounding unmasked words.
- Next Sentence Prediction (NSP) serves as another pre-training objective within the BERT framework. In this task, the model is presented with two consecutive sentences and tasked with classifying the sequential relationship between them, specifically whether the second sentence logically follows the first based on contextual coherence.
- Fine-tuning (not shown in the figure): After pre-training, the model is further adapted to a specific task (e.g., industry classification) by fine-tuning it on a labeled dataset relevant to that task. Fig. 2

This figure depicts the process of fine-tuning a pre-trained BERT model for industry classification. The pre-trained model parameters are adjusted using a dataset of labeled text specific to a particular industry. This fine-tuning step allows the model to learn the key features that distinguish different companies within that industry Fig. 3

3.2.3. BERT model architecture

- L = 12 (number of transformer blocks)
- H = 768 (hidden layer dimension)
- A = 12 (number of attention heads)
- Total number of parameters: 110M

3.3. Model training and validation

- Data Split: A dataset of 6056 industry-specific data points was used for training and validation. The data was split into three sets:
  - Training set (5056 data points)
  - Validation set (500 data points)
  - Test set (500 data points)
- Hyperparameters: During training, several hyperparameters were defined:
  - Maximum number of epochs: 10 (one epoch trains on the entire training set)
  - Batch size: 32 (number of training samples processed at once)
  - Early stopping: Training stops if the validation loss doesn't decrease for 1000 consecutive iterations.
- Evaluation: Model performance was evaluated every 50 iterations by monitoring the validation loss. The training process aimed to minimize the total loss on the validation set.

Fig. 4 shows the model training, validation, and testing results for the high-tech industry.

4. Empirical results and discussion

This section presents the results of our preliminary study. After model training, a large-scale test was conducted on 87,882 industry data points. The model correctly predicted 80,558 data points (represented by the sum of the main diagonal in a confusion matrix, Table 3). This translates to an overall accuracy of 92%. The remaining 7324 data points were misclassified (represented by the sum of the off-diagonal elements). The findings of this study substantiate the efficacy of the BERT-based model for the designated industry classification task. These results contribute to the growing body of evidence supporting the

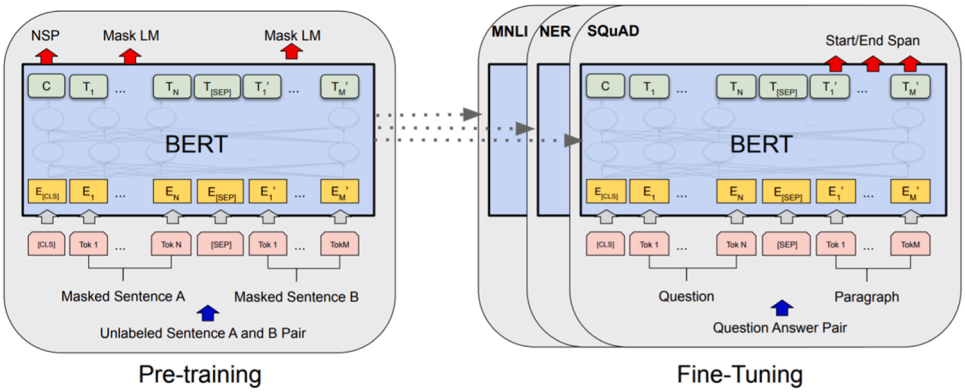


Fig. 1. BERT's Pre-training and Fine-Tuning Process.

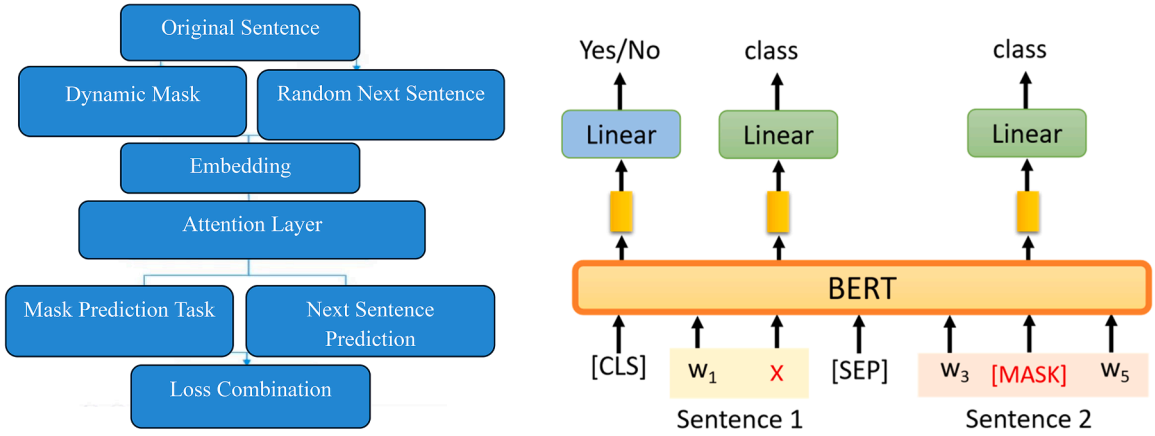


Fig. 2. Fine-tuning the BERT Model for Industry Classification.

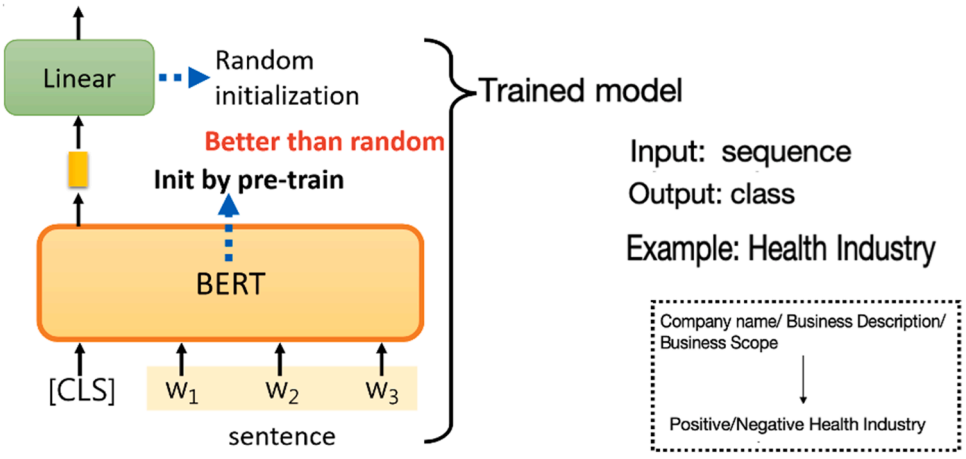


Fig. 3. The process of applying the BERT model to Industry Classification.

```
Epoch [8/10]
Iter:    1150,  Train Loss:  0.28,  Train Acc: 90.62%,  Val Loss:  0.15,  Val Acc: 94.80%,  Time: 12:17:32
Iter:    1200,  Train Loss: 0.038,  Train Acc: 100.00%,  Val Loss: 0.078,  Val Acc: 97.80%,  Time: 12:51:16 *
Iter:    1250,  Train Loss: 0.035,  Train Acc: 100.00%,  Val Loss:  0.12,  Val Acc: 95.20%,  Time: 13:25:07
Epoch [9/10]
Iter:    1300,  Train Loss:  0.14,  Train Acc: 96.88%,  Val Loss:  0.11,  Val Acc: 96.00%,  Time: 13:58:44
Iter:    1350,  Train Loss: 0.071,  Train Acc: 96.88%,  Val Loss: 0.098,  Val Acc: 97.00%,  Time: 14:32:16
Iter:    1400,  Train Loss:  0.31,  Train Acc: 90.62%,  Val Loss: 0.086,  Val Acc: 97.40%,  Time: 15:06:04
Epoch [10/10]
Iter:    1450,  Train Loss:  0.33,  Train Acc: 90.62%,  Val Loss:  0.12,  Val Acc: 96.60%,  Time: 15:39:53
Iter:    1500,  Train Loss: 0.065,  Train Acc: 96.88%,  Val Loss:  0.13,  Val Acc: 96.00%,  Time: 16:13:27
Iter:    1550,  Train Loss: 0.019,  Train Acc: 100.00%,  Val Loss: 0.087,  Val Acc: 97.40%,  Time: 16:47:21
Test Loss:  0.14,  Test Acc: 95.40%
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

Notbelonging     0.9710    0.9360    0.9532       250
   belonging     0.9382    0.9720    0.9548       250

    accuracy                         0.9540       500
   macro avg     0.9546    0.9540    0.9540       500
weighted avg     0.9546    0.9540    0.9540       500

Confusion Matrix...
[[234  16]
 [  7 243]]
Time usage: 0:03:06
```

Fig. 4. Model Training, Testing, and Validation Results in the High-Tech Industry.

Table 3
Training results for healthy industry.

	Model recognition It doesn't belong	Model recognition belong
Manual identification It doesn't belong	67,979	6750
Manual identification belong	574	12,579

potential of deep learning models in tackling intricate text classification problems.

4.1. Explanations

- Removed unnecessary repetition of the figure title.
- Clarified the meaning of "predicted correctly" and "predicted wrong" by referencing a confusion matrix (assuming Table 3 shows this).
- Improved sentence flow and conciseness.

It is worth noting that the model achieves high efficiency for NLP tasks. When tested on a large dataset using an Intel(R) Core(TM) i7–7700 CPU @ 3.60GHz PC, the model takes an average of 5 seconds to classify a single piece of company data.

The validation set results, shown in Table 4, reveal the model's ability to distinguish between various industry classes. The model achieves a high degree of precision in categorizing business descriptions, achieving scores between 84.11% and 99.66% across the 16 industry categories. Our findings are consistent with previous research by Santiago González [17] who reported an accuracy of 83.64% on their BERT industry classification task, and Jagrić et al. [16] reported a classification accuracy range of 83.5% to 92.6% across thirteen industry categories.

The model achieved particularly high accuracy in classes like modern light industry and textiles (99.66%) and green food (99.66%). This indicates that the descriptions within these categories likely exhibit unique characteristics. such as specific terminology or sentence structure, that the model can effectively learn and leverage for accurate predictions. The classification accuracy for all industry categories except "new display and intelligent terminal" ranged from 84.11% to 99.11%. The latter category achieved an accuracy slightly above 84.11%.

This showcases a balanced and resilient performance across a varied array of classes. Sustaining this equilibrium is vital in multiclass classification endeavors to prevent the model from exhibiting bias towards particular classes, a hurdle frequently tackled through methodologies such as class weighting. The validation process yielded promising findings, substantiating the model's capacity for effective learning and generalization across a diverse spectrum of industry classifications.

5. Conclusion

This study investigated the application of Natural Language Processing (NLP) and machine learning techniques for text-based industry classification. We leveraged a pre-trained BERT model to develop a system that groups companies based on their business descriptions. This work demonstrates that the BERT-based classification system achieves superior performance compared to traditional methods in some metrics, such as accuracy. The overall recognition rate ranged from 84.11% to 99.66% across 16 industry categories, highlighting the model's effectiveness. The model's exceptional accuracy in specific classes indicates its capability to grasp distinct industry traits. Furthermore, the model's consistent achievement of moderate to high accuracy across diverse industry classes underscores its robust generalizability to a wide range of textual data.

The encouraging results from this study suggest that the BERT model, when subjected to meticulous fine-tuning, has the potential to emerge as a valuable tool for automated industry classification tasks.

Table 4
Test results of industry classes with number of total cases and accuracy.

Industrial types	Industrial classification predicated	Total enterprise cases	Test accuracy rate
Six major strategies pillar industries	Equipment Manufacturing	66,638	98.72%
	Electronics Manufacturing	43,631	93.10%
	New Type of Building Materials	73,141	95.41%
	Advanced Metal Materials	121,318	90.48%
	Modern Light Industry and Textiles	76,219	99.66%
	Green Food	110,651	99.66%
Ten major strategies emerging industries	Biopharmaceuticals	47,743	94.22%
	Cybersecurity	70,219	96.99%
	New Energy and Networked Automobiles	105,566	86.59%
	New Display and Intelligent Terminal	97,982	84.11%
	5G	61,984	91.18%
	New Generation Artificial Intelligence	25,114	92.92%
	Intelligent Sensors	88,155	95.17%
	Energy Saving and Environmental Protection	63,314	94.52%
	Intelligent Equipment	90,084	93.93%
	Nylon New Materials	106,091	99.11%

Future research endeavors could explore further refinements in this domain, including:

1. Enhancing BERT with Supervised Learning: Explore how incorporating supervised learning approaches can further improve the model's performance.
2. Industry Classification Over Time: Investigate the use of company annual reports to analyze changes in industry classification over time.
3. Enriched Data Sources: Explore incorporating additional data sources, such as supply chain overlap and broker research, which might provide complementary information for classification.

By exploring these directions, we can further refine and enhance the effectiveness of NLP-based industry classification models.

CRediT authorship contribution statement

Baocheng Yang: Project administration. **Bing Zhang:** Conceptualization, Data curation, Methodology, Writing – review & editing. **Kevin Cutsforth:** Data curation, Formal analysis, Writing – original draft. **Shanfu Yu:** Project administration. **Xiaowen Yu:** Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] G. Hoberg, G.M. Phillips, Text-based network industries and endogenous product differentiation, J. Politi. Econ. 124 (5) (2016) 1423–1465.
[2] C. Lee, P. Ma, C. Wang, Search-based peer firms: Aggregating investor perceptions through internet co-searches, J. Financ. Econom. 116 (2) (2015) 410–431.

[3] M. Dalziel, X. Yang, S. Breslav, A. Khan, J. Luo, Can we design an industry classification system that reflects industry architecture? J. Enterprise Transform. (2018) 1–25, https://doi.org/10.1080/19488289.2017.1419319.
[4] J. Oyarzun, The imitation game: an overview of a machine learning approach to code the industrial classification, in: Proceedings of Statistics Canada Symposium 2018 Combine to Conquer: Innovations in the Use of Multiple Sources of Data, 2018.
[5] R. Dolphin, B. Smyth, R. Dong, A machine learning approach to industry classification in financial markets, in: Irish Conference on Artificial Intelligence and Cognitive Science, Springer Nature Switzerland, Cham, 2022, pp. 81–94.
[6] D. Kim, H.-G. Kang, K. Bae, S. Jeon, An artificial intelligence-enabled industry classification and its interpretation, Internet Res. 32 (2) (2022) 406–424, https://doi.org/10.1108/INTR-05-2020-0299.
[7] Rizinski, M., Jankov, A., Sankaradas, V., Pinsky, E., Mishkovski, I., & Trajanov, D. (2023). Company classification using zero-shot learning. *ArXiv, abs/2305.01028*.
[8] G. Bonne, A.W. Lo, A. Prabhakaran, K.W. Siah, M. Singh, X. Wang, P. Zangari, H. Zhang, An artificial intelligence-based industry peer grouping system, J. Financ. Data Sci. 4 (2) (2022) 9–36.
[9] A. Tagarev, N. Tulechki, S. Boytcheva, Comparison of machine learning approaches for industry classification based on textual descriptions of companies, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2–4 September 2019, Varna, Bulgaria, 2019, pp. 1169–1175.
[10] S. Slavov, A. Tagarev, N. Tulechki, S. Boytcheva. Company Industry Classification with Neural and Attention-Based Learning Models, 2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2019, pp. 1–7, https://doi.org/10.1109/BdKCSE48644.2019.9010667.
[11] H. Gao, J. He, K. Chen, Big Data & Innovative Financial Technologies Research Paper Series, 2020.
[12] S. Wang, Y. Pan, Z. Xu, B. Hu, X. Wang, Enriching BERT with knowledge graph embedding for industry classification. Advances in Knowledge Discovery and Information Retrieval, Springer, 2021, https://doi.org/10.1007/978-3-030-92310-5_82 (pp. xxx-xxx).
[13] L. Xu, B. Ji, Industry Classification Algorithm Based on Improved BERT Model. https://doi.org/10.1145/3573428.3573743.
[14] T. Ito, J. Camacho Collados, H. Sakaji, S. Schockaert, Learning company embeddings from annual reports for fine-grained industry characterization, in: Proceedings of the Second Workshop on Financial Technology and Natural Language Processing (IJCAI), January, 2021.
[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.
[16] T. Jagrič, A. Herman, AI model for industry classification based on website data, Information 15 (2) (2024) 89, https://doi.org/10.3390/info15020089.
[17] S. González-Carvajal, E. Garrido-Merchán (2020). Comparing BERT against traditional machine learning text classification. *ArXiv, abs/2005.13012*.