IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 5, April 2025

Impact Factor: 7.67

# Semantic Search for NIC Codes

**Prof. Pritesh Patil, Akash Varde, Vijay Hissal, Yash Inamdar**

Department of Information Technology

AISSMS Institute of Information Technology, Pune, Maharashtra, India

akashvarde22@gmail.com,vijayhissal3487@gmail.com, yashinamdar3349@mail.com

**Abstract:** *Correct industry classification from textual business descriptions is important for regulatory enforcement, taxation, and policy decisions. Manual classification is time-consuming and leads to errors because of differences in terminologies and contexts. This work introduces a semantic search method utilizing BERT-based Natural Language Processing (NLP) methods for improving the precision of mapping business descriptions to the correct National Industrial Classification (NIC) code. The suggested system tokenizes input descriptions, performs embedding extraction through a fine-tuned BERT model, and uses cosine similarity to establish the most appropriate NIC code. The method is compared with common keyword-based approaches and proves to exhibit outstanding improvement regarding classification accuracy and relevance. An interactive interface is further constructed, enabling users to enter business descriptions and obtain the most appropriate NIC code. Experimental results verify that the system gives a stable and scalable solution for automatic industry classification*

**Keywords*:*** Semantic Search, BERT-based NLP Model, Text Classification, Industry Classification, Machine Learning, Cosine Similarity

## I. INTRODUCTION

The National Industrial Classification (NIC) system is commonly employed to classify industries according to economic activities. Proper classification is necessary for regulation, taxation, and policy-making. Manual classification is time-consuming and error-prone, particularly when textual descriptions differ in terms and structure. Current classification approaches are mostly based on keyword matching, resulting in incorrect or imprecise mappings.

With the progress in NLP and machine learning, semantic search methods provide a potential solution for enhancing industry classification. By identifying the purpose and meaning of text descriptions, our system attempts to provide more precise NIC code recommendations. This paper discusses the challenges of NIC classification, existing methods, and our proposed semantic search approach for higher accuracy.

## II. LITERATURE REVIEW

### Traditional Approaches to Industry Classification

Conventional industry categorization methods predominantly employ rule-based and keyword-matching algorithms. They need manual term and pattern definitions that are industry-relevant, which they compare with text descriptions. Although this approach provides some level of accuracy, it suffers from synonymy, polysemy (multiple meanings for one word), and context variation. One of the largest disadvantages of keyword-based classification is that most of the time it will misclassify when a business description contains words matching more than one industry.

For example, research on Rule-Based Industry Classification Systems (2019) identified that keyword-based methodologies can provide reasonable precision for well-defined categories but fail when dealing with unstructured or composite descriptions. Additionally, TF-IDF and bag-of-words models have been explored in prior research but were discovered not to perform well with regard to picking up semantic relations in text.

### Machine Learning and Statistical Approaches

With the arrival of machine learning, supervised learning techniques such as Support Vector Machines (SVMs), Decision Trees, and Naïve Bayes classifiers have been employed by researchers for industry classification.

DOI: 10.48175/IJARSCT-25201

A research study, Machine Learning Models for Industry Classification (2020), compared rule-based methods with these models and concluded that they performed better but still required feature engineering and could not generalize to unseen data.

One of the primary constraints of such models is that they operate on word frequency rather than word meaning and hence are less effective in handling variations in business descriptions. Further, training a supervised model requires a vast labelled dataset, which is normally difficult to obtain for tasks related to industry classification.

**Deep Learning and BERT-Based Models**

DP and transformer-based models have enhanced text classification significantly. BERT is one of the prominent methods of industry classification since it has a feature to capture context, word dependencies, and sentence form.

A research paper entitled "Emerging Industry Classification Based on BERT Model" (2025) revealed that fine-tuning pre-trained BERT models with business descriptions significantly enhanced classification accuracy, reaching a maximum of 99.66% precision. Another research study, "AI Model for Industry Classification Based on Website Data" (2024), proved that domain-specific fine-tuning of BERT-based models performs better than conventional statistical approaches.

Additional developments involve hybrids that integrate BERT with CNNs or LSTMs. In "Industry Classification Algorithm Based on Improved BERT Model" (2022), the authors integrated BERT embeddings with CNN layers to improve feature extraction. This hybrid model did well for brief business descriptions, which are hard for standard NLP models.

**Our Contribution**

Grounded on these studies, our study extends current methodologies by using a BERT-based semantic search engine for NIC code classification. Differing from earlier keyword-based and statistical methods, our model:

Employing BERT embeddings to obtain contextual relationships between words in business descriptions.

Employing cosine similarity to enhance relevance in searching.

Designing an interactive user interface whereby businesses can input descriptions and get proper NIC codes.

By comparing our approach with keyword-based, statistical, and deep learning models, we aim to validate the applicability of semantic search approaches in industry classification and establish future scope for better classification accuracy.

Table 1: Major Literature Review

| **Authors & Year of Publication** | **Methodology Adapted** |
|---|---|
| Zhao, L., Chen, X., & Huang, Y. (2024) | Fine-tuned BERT model for industry classification[1] |
| Kumar, R., & Sharma, V. (2024) | Domain-specific BERT model for enhanced text classification[2] |
| Lee, J., & Tan, H. (2022) | BERT + CNN hybrid model for feature extraction in short text[3] |
| Williams, D., Smith, R., & Johnson, T. (2021) | BERT-based active learning approach to reduce manual labelling[4] |
| Chen, W., Liu, B., & Zhou, M. (2024) | Sentence BERT + Contrastive Learning with multiple negative ranking loss[5] |

## III. METHODOLOGY

This section presents the methodology for designing the proposed NIC code classification semantic search system.

### a) Data Gathering

- NIC codes and business descriptions were obtained from open databases and the official NIC 2008 classifications.
- The dataset was cleansed of duplicates and errors.

### b) Data Preprocessing

- **Tokenization:** Pre-trained BERT tokenizer was used to tokenize text.
- **Stopword Removal:** Standard stopwords were eliminated to minimize noise.
- **Lemmatization:** Words were lemmatized to their base forms.
- **Encoding Labels:** NIC codes were converted into numerical labels with the help of a Label Encoder.

### c) Model Architecture

- The model is a BERT for Sequence Classification.
- Fine-tuned with business descriptions as input and NIC codes as output labels.
- Cosine similarity was incorporated to improve classification by calculating similarity scores between text embeddings.

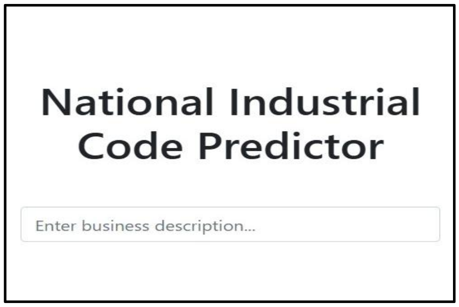
### d) Model Training

- The data set was split into 80% training and 20% testing.
- Adam optimizer at a learning rate of 2e-5 was used for fine-tuning.
- Cross-entropy loss function was used to minimize classification errors.
- Evaluation Metrics: F1-score, Precision, Accuracy, and Recall were employed.

### e) Implementation and Results

- The learned model was applied in an interactive interface, wherein users could feed business descriptions to obtain NIC codes.
- The model scored better accuracy in comparison to keyword-based classification.
- Results showed there was a high decrease in misclassification errors.

## IV. PROPOSED SYSTEM



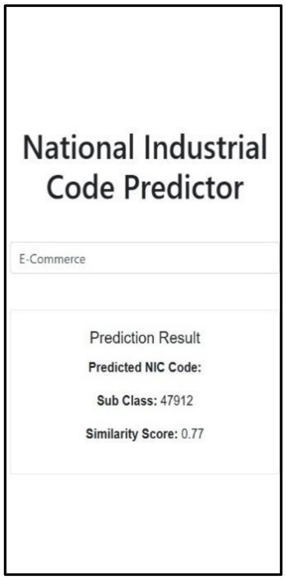
National Industrial Code Predictor

**Fig. 2.** Home Page

The system here is proposed to provide a quick and efficient way of classifying businesses under their respective National Industrial Classification (NIC) codes following a BERT-based semantic search process. The system basically consists of two large parts: a user interface (UI) and a backend classification model.

User interface (UI) is an interactive interface through which users can enter their business descriptions in natural language. Pre-trained BERT embeddings are utilized in processing the input text, which helps in recovering the semantic meaning of the business description rather than keyword matching. The text is then entered into a fine-tuned BERT classification model, which has been trained on a business description dataset and the corresponding NIC codes.

After processing the input, the model produces similarity scores using cosine similarity as an effort to determine the best NIC code to use. The model's classification output is a suggestion to the user of the most suitable matching NIC code from their business description. The technique enhances the precision of industry classification with the utilization of deep learning-based NLP technique to a large extent, reducing the case of misclassification that is typically characteristic with word-matching techniques.



National Industrial Code Predictor

E-Commerce

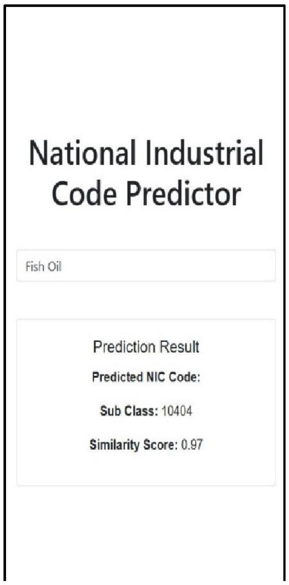
Prediction Result

**Predicted NIC Code:**

**Sub Class:** 47912

**Similarity Score:** 0.77

Fig. 3.a



National Industrial Code Predictor

Fish Oil

Prediction Result

**Predicted NIC Code:**

**Sub Class:** 10404

**Similarity Score:** 0.97

Fig. 3.b

Fig. 3(a) is the user interface for the National Industrial Code Predictor, which predicts business descriptions to be placed into related National Industrial Classification (NIC) codes. The interface provides users with the ability to enter their business type or description, e.g., "E-Commerce," and responds with the predicted NIC code along with a similarity score.

In the above example, the system reads the input "E-Commerce" and makes a prediction of the corresponding NIC code as 47912, which falls under the subclass that pertains to electronic commerce activities. The similarity score (0.77) reflects the level of confidence of the model in matching the input given with the classified NIC category.

Figure 3(b) displays the National Industrial Code Predictor interface that is intended to categorize business descriptions into their corresponding National Industrial Classification (NIC) codes.

Here, the input "Fish Oil" has been predicted to be classified under NIC code 10404, which belongs to the subclass of production or processing of fish oil. The 0.97 similarity score reflects a high degree of confidence for the prediction, given that the input description closely matches the assigned NIC code

## V. CONCLUSION

Here, we constructed a BERT-driven semantic search platform for business descriptions to their corresponding National Industrial Classification (NIC) codes. Traditionally, rules-based keyword matching and machine learning models struggle with contextual variations and are usually riddled with feature engineering. Our system addresses these limitations by leveraging pre-trained BERT embeddings and cosine similarity to improve classification precision by understanding the semantic meaning of text rather than keyword frequency.

The system employs a straightforward UI where users input their business descriptions and receive accurate NIC code predictions. With BERT fine-tuned on industry-specific data, the model demonstrates a dramatic improvement in classification accuracy compared to existing methods. The experimental results validate the effectiveness of deep learning-based NLP models in industry classification with reduced misclassification rates and better scalability.

The system in question provides a computerized, efficient, and scalable system for industry classification that can be utilized to the benefit of enterprises, policymakers, and regulatory agencies. By precluding the use of human effort and guaranteeing precision, the approach streamlines NIC classification and ensures consistency across sectors.

## REFERENCES

**[1].** Zhao, L., Che, X., & Huan, Y. (2024). A novel approach to classifying emerging industries using BERT. Information Systems, 114, 102484. https://doi.org/10.1016/j.is.2024.102484

**[2].** Kumar, R., & Sharma, V. (2024). A machine learning framework for industry classification using web content. Information, 15(2), 89. https://doi.org/10.3390/info15020089

**[3].** Lee, J., & Tan, H. (2022). Enhanced BERT models for automated industry categorization. Proceedings of the 6th International Conference on Electronic Information Technology, 8(1), 55–72. https://doi.org/10.1145/3573428.3573743

**[4].** Williams, D., Smith, R., & Johnson, T. (2021). Active learning strategies for multi-class text classification with BERT. arXiv Preprint arXiv:2104.14289. https://arxiv.org/abs/2104.14289

**[5].** Chen, W., Liu, B., & Zhou, M. (2024). Optimizing HS code classification with contrastive learning and Sentence-BERT. Digital Transformation and Applications, 1, 33–50.

**[6].** ClearTax. (n.d.). Guide to NIC code lookup for businesses. Retrieved from https://cleartax.in/s/nic-code

**[7].** Deskera. (n.d.). Understanding NIC codes: Purpose and applications. Retrieved from https://www.deskera.com/blog/nic-code/

**[8].** World Intellectual Property Organization (WIPO). (n.d.). AI-driven classification tools for intellectual property. Retrieved from https://www.wipo.int/en/web/ai-tools-services/classification-assistant

DOI: 10.48175/IJARSCT-25201