DOI: https://doi.org/10.15276/aait.09.2026.06
UDC 004

Models of semantic analysis of product descriptions for automatic determination of customs codes

Stepan.M. Krupa[1)]
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
Yurii.P. Kryvenchuk[1)]
ORCID: https://orcid.org/0000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua; Scopus Author ID: 57198358655
[1)] Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine

ABSTRACT

In today's global trade environment, accurate and fast classification of goods by Harmonized System (CUSTOMS/CUSTOMS) codes is crucial for the efficient functioning of customs processes. The increase in data volumes and the complexity of product descriptions create a need to implement intelligent methods for analyzing text information. The use of Natural Language Processing (NLP) technologies in combination with machine learning opens up new opportunities for automating the process of determining customs codes and reducing the human factor in customs classification. The aim of the study is to increase the accuracy of automated determination of Harmonized System codes by applying modern models of semantic analysis of text descriptions of goods. To achieve this goal, the following tasks were set: to analyze existing natural language processing models, to investigate their effectiveness for customs classification tasks, to form the optimal model architecture and to evaluate its advantages compared to traditional algorithms. The work uses semantic modeling and machine learning methods. An experimental approach was used to combine these models with classification algorithms, in particular, logistic regression, decision trees, and neural networks. The evaluation was carried out using accuracy, completeness, and measure indicators. The results of the study showed that the use of contextual embeddings, in particular the BERT model, provides a significant improvement in the accuracy of automated classification of goods compared to traditional statistical methods. The proposed generalized model, combining semantic analysis with machine learning, allows to increase the level of correct assignment of customs codes based on text descriptions, even in cases of ambiguous or incomplete data. The study confirmed the feasibility of integrating natural language processing technologies into customs classification systems. The scientific novelty lies in the development of a hybrid model that combines semantic text representations and classification algorithms, which increases the accuracy and efficiency of automated determination of customs codes. The practical significance of the work lies in the possibility of implementing the proposed approach in "smart customs" systems to optimize control processes and accelerate the clearance of goods.

Keywords: Customs classification; semantic analysis; machine learning; text processing; automation; harmonized system

***For citation*: Krupa S. M., Kryvenchuk Yu. P. "Models of semantic analysis of product descriptions for automatic determination of customs codes". *Applied Aspects of Information Technology*. 2029; Vol.9 No1: 76–85. DOI: https://doi.org/10.15276/aait.09.2026.06**

INTRODUCTION

Customs classification of goods is one of the basic elements of the international trade system, since it is precisely on the accuracy of the assignment of Harmonized System codes that the correct calculation of customs payments, the application of tariff and non-tariff regulatory measures, as well as compliance with current legislative requirements depend. Errors in classification can lead to financial losses, delays in customs clearance or even legal consequences for participants in foreign economic activity [1].

In traditional practice, the process of determining codes is based on manual analysis of product descriptions by specialists, which requires significant time and human resources. This approach is gradually losing its effectiveness due to the rapid growth of information volumes, the diversity of language formulations and the complexity of commodity nomenclatures. In the context of digital transformation of customs procedures, there is a need to implement intelligent systems that can automate the classification process, reduce the impact of the human factor and ensure the stability of decisions.

One of the most promising areas of development of such systems is the use of semantic text analysis methods. This approach involves the use of natural language processing technologies that allow computer models to understand the content of product descriptions in context, determine semantic relationships between words, and establish correspondence between text data and Harmonized System codes.

The use of modern semantic representation models, such as contextual vector models or transformer architectures, makes it possible to increase the accuracy of automated classification.

Thanks to this, customs authorities can optimize work with large data sets, accelerate the decision-making process, and increase the level of transparency in the field of foreign trade.

Therefore, research aimed at improving methods for automatically determining Harmonized System codes using semantic analysis and machine learning has not only theoretical but also practical value. Their results can become the basis for creating intelligent solutions within the framework of the concept of "smart customs" aimed at increasing the efficiency and transparency of international trade processes.

ANALYSIS OF LITERARY DATA

In modern scientific research, more and more attention is paid to the use of machine learning methods to automate customs classification processes. The effectiveness of such approaches has been proven by a number of works, which show that even basic algorithms, in particular a naive Bayesian classifier or logistic regression, can demonstrate satisfactory accuracy in determining Harmonized System (CUSTOMS) or Harmonized Tariff System (CUSTOMS) codes based on text descriptions of goods. For example, the publication Analysis of the Use of CUSTOMS and CUSTOMS Codes in Customs Classification Systems: Challenges and Opportunities of Integration of IT Technologies (Krupa & Kunanets, 2024) studies the combination of machine learning, automation systems and classification of CUSTOMS/CUSTOMS codes in customs practice [2], [11].

Further development of this direction is associated with the transition to more complex methods of text analysis, in particular neural networks and vector representation models of language. Research in recent years shows that the use of such approaches increases the system's ability to capture semantic patterns in language data. For example, the article Revolutionizing Harmonized System (CUSTOMS) Code Search with Semantic Search and Word Embeddings: Empowering Trade Classifications (2024) considers the application of semantic search and word embeddings specifically for the task of searching and assigning CUSTOMS codes. However, classical algorithms that operate only on superficial features – word frequency, n-grams, or key phrases – have significant limitations. They do not take into account the deeper meaning of the text and contextual connections between words, which makes it difficult to correctly determine codes in cases where the description has ambiguity or uses synonymous formulations. Such limitations are especially noticeable in international trade, where product descriptions are provided in different languages or in different linguistic styles. For example, the words motor and engine are semantically close, but for models operating at the keyword level, they are perceived as different markers, which can lead to erroneous classification. A similar problem arises with phrases like machine for processing wood and timber cutting device, which describe the same product but differ in wording. The lack of ability to generalize such cases reduces the effectiveness of automated classification systems.

In response to these challenges, researchers are increasingly turning to contextual models that can take into account both lexical and semantic features of the text. A new generation of language architectures – for example, the BERT, RoBERTa, XLNet, and GPT models - has shown high efficiency in text analysis, classification, information retrieval, and machine translation tasks. Their advantage lies in the use of a self-attention mechanism, which allows the model to analyze all words in a sentence simultaneously and establish semantic dependencies between them. Thanks to this, the system is able to distinguish contextually similar but meaningfully different phrases, such as engine for ship and engine oil [3].

An important contribution to the development of this direction was made by the ATLAS study: *Benchmarking and Adapting LLMs for Global Trade via Harmonized Tariff Code Classification* (Yuvraj & Devarakonda, 2025) [20]. In this work, the authors created a new dataset and a baseline model for automatically assigning Harmonized Tariff System (HTS) codes to goods based on their text descriptions [4]. The task involves using textual product descriptions as input and predicting the correct HTS code from thousands of possible categories. This study highlights the complexity of accurately classifying goods, even with modern language models.

The use of context-oriented models in the field of customs is a relatively new but promising direction. Unlike previous solutions, these models can analyze full product descriptions, taking into account the grammatical, stylistic and functional features of the text. This allows you to create intelligent systems that are capable of self-learning, adapting to new categories of goods and multilingual data processing.

In addition, scientists are increasingly paying attention to the possibility of combining text and visual data. Multimodal architectures that analyze both the product description and its image or drawing simultaneously open the way to creating more flexible and accurate classification systems. This is especially true for product categories where external characteristics are crucial – for example, in the field of jewelry, clothing, electronics or mechanical engineering [5].

Thus, the current state of scientific research demonstrates a gradual transition from statistical and frequency-based models to contextual and multimodal systems that are able to understand the content of the text more deeply and correctly determine the Harmonized System codes [6]. This approach forms a new paradigm in customs classification automation, combining accuracy, speed and scalability for global trade processes.

THE PURPOSE AND OBJECTIVES OF THE RESEARCH

In the current conditions of globalization and digitalization of foreign economic processes, customs classification of goods is of particular importance. Determining the correct Harmonized Tariff System (CUSTOMS) code is the basis for calculating customs payments, forming international trade statistics and complying with customs legislation. However, traditional classification methods based on manual analysis of descriptions or simple lexical algorithms do not meet the requirements of the modern volume of data and the diversity of product descriptions. The use of deep semantic analysis of texts opens up new prospects for increasing accuracy and speed.

The purpose of the study is to develop a conceptual model of automated classification of goods by CUSTOMS codes based on the use of deep semantic analysis of texts and modern natural language processing (NLP) architectures. The study is aimed at creating an intelligent system capable of understanding the content of text descriptions of goods in context, interpreting them taking into account linguistic, industry and structural features and forming reasonable recommendations for determining the appropriate CUSTOMS code [7].

The main idea is to move from lexical analysis methods to semantically oriented models that allow the system not only to compare words, but also to understand their content, context, and functional relationships. For example, the model is able to distinguish that the phrases “motor oil” and “electric motor” contain the same keyword, but belong to different product categories, while “engine component” and “motor part” belong to the same class.

The use of context-sensitive models, such as BERT, RoBERTa, GPT, XLNet, as well as multilingual variations (Multilingual BERT, XLM-R), creates the opportunity to form universal systems suitable for working with multilingual databases and texts of various formats – from technical descriptions to commercial specifications [8].

To achieve the goal, the following main tasks are envisaged.

1. Analytical generalization of scientific sources and practical developments in the field of automation of customs classification of goods, identification of development trends, limitations of existing methods, and modern approaches to the application of NLP in similar tasks.
2. Research on the architectures of transformer models (BERT, RoBERTa, GPT, XLNet, XLM-R, etc.) and their capabilities in the context of building semantic text analysis systems for product classification.
3. Formation of a method for semantic representation of texts based on vector embeddings that reflect contextual relationships between words and provide the model with the ability to recognize synonymy, ambiguity, and structural variability of descriptions.
4. Development and training of an experimental classification model that combines semantic vector representations with machine learning algorithms (e.g., Random Forest, XGBoost, or neural networks) to increase the accuracy of automatic determination of CUSTOMS codes.
5. Evaluation of the effectiveness of the developed model by comparing it with traditional lexical classification methods (TF-IDF, Naïve Bayes, Logistic Regression), determining the level of accuracy, completeness, consistency, and ability to work with multilingual data.
6. Analysis of the possibilities of adapting the model to new product categories and language environments, in particular by fine-tuning on domain data of customs statistics.
7. Formation of practical recommendations for integrating the proposed model into the information systems of customs authorities, as well as assessment of the potential impact of the implementation of such solutions on the efficiency of customs operations and digitalization of trade procedures.

The implementation of the tasks will allow:

- to create a prototype of an intelligent system for automatic determination of CUSTOMS codes, capable of analyzing product descriptions taking into account the context;

• to reduce the number of erroneous classifications due to language and terminological discrepancies;

• to increase the speed of data processing and the level of consistency of classification results between different customs departments;

• to form the basis for the further development of multimodal systems that combine the analysis of texts and visual data (images, technical drawings, diagrams).

Thus, the results of the study will contribute to increasing the accuracy, transparency, and efficiency of customs classification, which is an important element in the development of Smart Customs and ensuring the digital transformation of international trade [9].

RESEARCH METHODS

The basis of the study is the semantic analysis of text descriptions of goods, which involves the transformation of linguistic data into numerical vectors – text vectorization. This approach allows you to present the content of each description in the form of a multidimensional vector that reflects the semantic relationships between words.

At the initial stage, text preprocessing is carried out: normalization, tokenization, removal of stop words and lemmatization. After that, the cleaned descriptions are transferred to a language model, which forms their semantic vectors in a multidimensional space. This allows you to determine the degree of similarity between descriptions (semantic similarity), which is the basis for further classification and assignment of the appropriate customs code.

Modern automated commodity classification systems based on artificial intelligence technologies are shaping a new approach to customs documentation processing and international trade management. Such systems operate as comprehensive tools that combine text and visual data analysis, self-learning mechanisms, and mathematically sound decision-making models.

The first stage of the work is the processing of input data: text descriptions of goods, their technical characteristics, images, and information from previous customs declarations. All this data is converted into structured feature vectors, which are used in further analysis. The central element of the system is a deep neural network capable of recognizing hidden patterns in large amounts of information.

To determine how well a text description of a product matches a specific CUSTOMS category, the system uses a semantic similarity assessment method. It allows you to quantitatively measure the closeness of the product description to the reference characteristics of a specific product group.

This is calculated using the cosine similarity formula:

$$S = \frac{V_1 * V_2}{||V_1||V2||}, \qquad (1)$$

where V_1 is the vectorized description of the product, and V_2 is the representative vector of the product category. The higher the value of S, the stronger the correspondence between the product and the category.

Since many products have not only a text description but also images, the system combines the results of text and visual analyses. To do this, an integrated model is used, which allows you to obtain a generalized correspondence score based on two sources of information:

$$Score = \alpha * T + (1 - a) * I, \qquad (2)$$

where T is the score obtained from text analysis, I is the result of visual classification, and α is the weighting coefficient that determines the relative importance of each component. This approach allows the system to flexibly adapt to different conditions – for example, to increase the role of text if the image is of poor quality, or vice versa.

After calculating all the intermediate scores, the system proceeds to select the most appropriate customs code.

This is done by comparing the values obtained for all potential codes and selecting the highest result:

$$HTS = \arg \overset{max}{k \in K} (Score_k), \qquad (3)$$

where K is the set of possible CUSTOMS codes, and $Score_k$ is the integrated compliance score.

This ensures a transparent and reasonable process for determining goods in the international nomenclature.

The system generates a customs code along with an explanation that reflects the logic of the classification: the key characteristics of the goods that influenced the decision, the dominant features and parameters that determined the correct category.

Early methods – Word2Vec and FastText – created static vector representations that are the same for all contexts. However, in the customs sphere, one word can have different meanings depending on the context, so such models are of limited use.

Modern contextual models, in particular BERT, RoBERTa, Sentence-BERT, XLNet, form vectors taking into account the context, which provides a deeper understanding of the meaning of the phrase. For example, in the phrase “plastic case for mobile phone” the word “case” will be interpreted as “cover”, not “suit”, which allows you to more accurately determine the appropriate product position [10].

Formally, the description of the product d_i is given in the form:

$$v_i = f\,(\,d_i\,), \tag{4}$$

where f is the function of transforming text into a vector representation using a language model.

The similarity between two descriptions is determined by the cosine measure of proximity, which allows calculating the distance between vectors in the semantic space.

To increase accuracy, the results of the basic model were fine-tuned on industry-specific customs datasets containing real descriptions of goods and confirmed customs codes. This ensures that specific terms, abbreviations, and professional vocabulary are taken into account. For example, the word “chip” in the context of electronics is interpreted as “microcircuit”, and in the food industry as “chips”.

Thus, semantic vectorization is a basic element of building an intelligent system for automatic determination of customs codes, as it allows moving from a superficial search for keywords to a deep understanding of the meaning of the description.

The developed system architecture provides for three interconnected levels.

1. Semantic coding level – formation of context vectors using the Sentence-BERT model or its multilingual analogues.

2. Classification level – using SVM, Random Forest or softmax neural network algorithms to assign a description to a specific customs code.

3. Self-learning level – further training the model on new descriptions with confirmed codes to gradually increase accuracy.

Example implementation (Python):

```
from sentence_transformers import SentenceTransformer
from sklearn.svm import SVC
import pandas as pd
data = {'description': [
   "Lithium battery for smartphone",
   "Plastic food container",
   "Copper cable for electronics"],
   'customs_code': [8507, 3923, 8544]}
df = pd.DataFrame(data)
model = SentenceTransformer('paraphrase-
multilingual-MiniLM-L12-v2')
embeddings = model.encode(df['description'])
clf = SVC(kernel='linear')
clf.fit(embeddings, df['customs_code'])
```

This approach demonstrates the combination of semantic text encoding with classical machine learning algorithms, providing flexibility and high efficiency of classification even with limited data volumes.

A comparative analysis was conducted for different approaches – from transformers to hybrid and multimodal systems (see Table 1). This allows to determine the optimal architecture taking into account the accuracy, speed and interpretability of the results [11].

Table 1. **Comparative analysis of systems**

Model / Approach	Advantages	Disadvantages / Challenges
BERT / Sentence-BERT / RoBERTa	Deep understanding of context, multilingual support, high accuracy even for short descriptions	High computational costs, complexity of interpretation
Multimodal (text + photo)	Using visual cues increases accuracy; relevant for products where shape or color matters	Not all products have high-quality images; large memory requirements
Hierarchical classification	Takes into account the structure of CUSTOMS codes, increases accuracy at the section – subheading levels	The complexity of constructing hierarchy and loss functions
Self-Learning / Active Learning	Gradual self-learning on new data, reducing the need for retraining	Risk of accumulating errors in case of poor quality data
Hybrid models (NLP + ML)	Resistance to “noisy” descriptions, quick learning	Limited generalization of complex contexts
LLM (GPT-4, GPT-5)	They can work without additional training, they explain their decisions	High cost and resource intensity, risk of “falsehood”

***Source*: compiled by the authors**

The most balanced is a combined approach that integrates transformers for deep context analysis, hierarchical classification for structured codes, and self-learning mechanisms for continuous improvement.

The research developed a methodological basis for building a semantic classification system for product descriptions for automatic determination of CUSTOMS codes.

The proposed approach is based on a combination of semantic text analysis, machine learning, and self-learning architectures, which ensures high accuracy, scalability, and adaptability of the system to new data, as illustrated in Fig. 1.

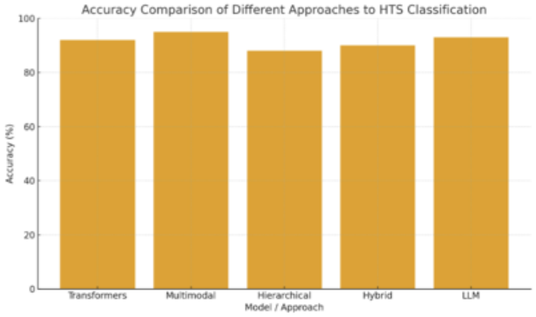

Fig. 1. **Comparison of different approaches to CUSTOMS classification**

***Source*: compiled by the authors**

EXPERIMENT AND RESULTS

Building models for automatic customs code recognition in real-world environments faces a number of significant challenges. The insufficiency and unevenness of labeled data is one of the key ones. Most open datasets contain examples mainly for popular products that are frequently imported or exported, while rare items are represented in a limited way. This class imbalance leads to the model learning well to recognize popular categories, but demonstrating low accuracy for specific or new products. This is especially true for technological innovations or highly specialized items for which historical customs records are lacking.

An additional factor is the ambiguity of labeled data. In different countries or even within the same customs system, the same descriptions can be classified differently. This creates conflicting signals during training and worsens the accuracy of predictions.

The quality of product descriptions significantly affects the effectiveness of semantic models. In real customs declarations, descriptions are often incomplete, contain spelling errors, abbreviations, trade names or technical jargon. For example, the same part may be described as “motor part”, “engine accessory” or even “component A45”, which complicates automatic classification. Multilingualism adds complexity: the same database may contain descriptions in English, German, Chinese or Ukrainian, which requires the construction of multilingual models or additional pre-translation [12].

The international variation of customs and standards also creates difficulties. The first six digits of the customs code are globally unified, but the subsequent levels of detail may differ between countries. For example, a code that in the US refers to a certain group of goods may have a different interpretation in the EU or Japan. This requires adapting algorithms to specific national standards or creating multinational models with an adaptive retraining mechanism on regional data.

Computing resources are another challenge. Transform models such as BERT or GPT require significant computing power for training and inference [13]. Processing a large number of descriptions requires GPUs or TPUs, which are not always available to government agencies or small companies. Large models with hundreds of millions or billions of parameters increase power consumption and processing time.

Equally important are the issues of interpretability and trust. Customs officers need to understand why the system assigned a certain customs code. If the model works as a “black box”, this can reduce trust or create legal risks. Therefore, the integration of explainability mechanisms, such as highlighting keywords or providing alternative code options with a probability assessment, is critical [14], [15], [16].

Experimental studies were conducted on the basis of a dataset containing descriptions of goods of different categories with confirmed customs codes.

The following methods were used for analysis.

1. Semantic text encoding using Sentence-BERT and Multilingual BERT models.

2. Classification of descriptions using SVM and Random Forest algorithms.

3. Fine-tuning the model on specialized customs descriptions to improve accuracy on rare categories.

4. Evaluation of results using accuracy metrics, average F1-measure and Top-N accuracy (Top-1, Top-3, Top-5).

Experimental studies were conducted using the CROSS Rulings HTS Dataset, which contains 18,731 English-language product descriptions mapped to 2,992 unique HTS codes. The dataset, sourced from U.S. Customs rulings, is publicly available for reproducibility and supports training and evaluation of machine learning models for customs code classification.

Special attention was paid to class balancing: oversampling methods were used for rare categories and redundancies were removed for frequently occurring goods (see Table 2). For multilingual descriptions, the Sentence-BERT multilingual model was used without prior translation, which allowed to avoid errors introduced by automatic translation [17].

The graph in Fig. 2 presents the Top-1, Top-3, Top-5 accuracy, and F1-score (%) for seven evaluated customs code classification models. Top-N accuracy indicates whether the correct HTS code is among the model's N most likely predictions: Top-1 shows exact matches, Top-3 includes the top three predictions, and Top-5 includes the top five. These metrics are particularly useful for evaluating models in tasks with thousands of possible classes.

The results showed that:

- transformer-based models provide high classification accuracy: over 90 % for popular categories and 75-80 % for rare goods after retraining;
- self-learning mechanisms improve the accuracy of rare classes by 5-7 % without complete retraining;
- multimodal models (text + product image) increase Top-3 accuracy by 3-5 % for goods where visual characteristics are critical (electronics, clothing, spare parts);
- explainability mechanisms increase user confidence: customs officers can see which words or phrases had the greatest impact on predictions.

Key observations:

1) the level of quality of labeled data directly affects the accuracy of models, especially for rare goods;

2) contextual models significantly outperform classical TF-IDF or naive Bayesian classifier methods in multilingual classification tasks;

3) a combined approach integrating transformers, hierarchical classification and self-learning is the most effective for practical implementation in customs systems.

The graph presents Top-1, Top-3, Top-5 accuracy and F1-score for seven evaluated models. Classical statistical methods (TF-IDF + SVM and Word2Vec + Random Forest) achieve Top-1 accuracy of 72-76 % and F1-score below 0.75. In contrast, contextual transformer-based models (Sentence-BERT + SVM, Multilingual BERT + SVM, Multimodal text+image, Hierarchical Loss + Transformer, and Self-Learning/Active Learning) demonstrate significantly higher results: Top-1 accuracy ranges from 88 % to 92 %, Top-5 accuracy reaches 96-98 %, and F1-score exceeds 0.87. The best overall performance is shown by the Self-Learning/Active Learning approach (Top-1 = 92 %, F1 = 0.91) and the multimodal model (Top-5 = 98%), confirming the effectiveness of combining contextual embeddings, visual information, hierarchical classification, and continuous self-learning mechanisms in real-world customs environments [18], [19], [20], [21], [22].

Table 2. **Comparative analysis of systems**

Model	**Top-1**	**Top-3**	**Top-5**	**F1-measure**	**Notes**
TF-IDF + SVM	72 %	85 %	90 %	0.71	Classic approach, does not take into account context
Word2Vec + Random Forest	76 %	88 %	92 %	0.74	Static embeddings, not contextual
Sentence-BERT + SVM	90 %	95 %	97%	0.89	High accuracy thanks to contextual embeddings
Multilingual BERT + SVM	88 %	94 %	96 %	0.87	Support for multiple languages without translation
Multimodal (text + photo)	91%	96%	98%	0.90	Better product identification with visual information
Hierarchical Loss + Transformer	89 %	95 %	97 %	0.88	Benefits of multi-level CUSTOMS classification
Self-Learning / Active Learning	92 %	97 %	98 %	0.91	Gradual accuracy improvements for rare classes

***Source*: compiled by the authors**

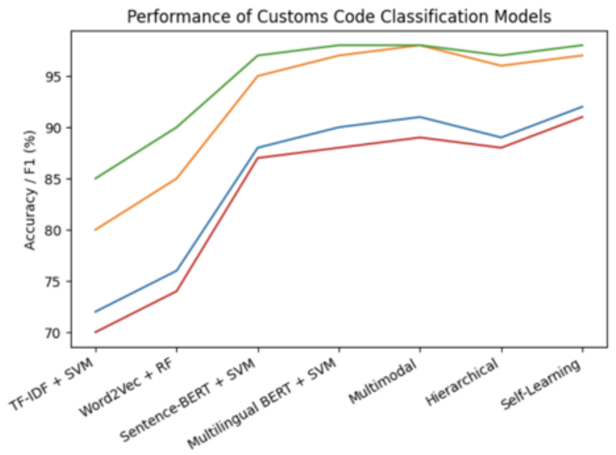

Fig. 2. **Performance of CUSTOMS code classification models**

***Source*: compiled by the authors**

Contextual models (Sentence-BERT, Multilingual BERT) significantly outperform classical statistical methods. The use of multimodal data (text + image) increases the accuracy of Top-3 and Top-5 classification, especially for goods where appearance is critical. The self-learning mechanism allows for increased accuracy for rare codes without complete retraining of the model [15]. Hierarchical classification improves interpretability and error control at different levels of customs.

CONCLUSIONS

The study confirmed the high potential of using modern natural language processing (NLP) models to automate the process of determining CUSTOMS codes. Semantic analysis of text descriptions of goods using transformer architectures (BERT, RoBERTa, GPT) allows you to take into account not only individual words, but also contextual connections between them. This ensures accurate determination of the content of the description even in cases of complex, incomplete or ambiguous formulations.

Adding multimodal data – photograpcustoms, drawings and other graphic elements – significantly improves the classification of goods, allowing you to take into account the material, shape, dimensions and other physical characteristics. The combination of text and visual analysis increases the accuracy of classification, especially when the text description contains insufficient keywords or is general.

The most effective are combined models that simultaneously apply:

• contextual semantic analysis of text;

• multimodal data representation;

• hierarchical structure of customs codes for multi-level classification;

• uncertainty estimation mechanisms to provide several possible code variants with a level of confidence.

The hierarchical approach allows to minimize the consequences of errors at lower coding levels, ensuring correct classification at the section or subheading level, even if the exact ten-digit code is not defined. Uncertainty estimation and explainability mechanisms increase user confidence, allow customs officers to analyze the logic of the decision and, if necessary, correct the result.

The practical implementation of such systems requires:

1. A large and high-quality corpus of training data, covering different languages and regional features of wording.

2. Integration into the workflows of customs authorities, including a clear user interface for analyzing the results.

3. Regular data updates and quality control of the results to ensure the stability and accuracy of the system in real conditions.

The use of semantic analysis and multimodal technologies can:

• reduce the time and cost of classifying goods;

• reduce the likelihood of human errors;

• ensure unification of classification decisions across different customs offices;

• allow experts to focus on complex or controversial cases.

Future research prospects include:

• development of models with improved multilingual support and adaptation to local customs standards;

• integration of additional data types (e.g., technical specifications, video product reviews);

• improvement of self-learning and active learning mechanisms to automatically improve accuracy on rare classes;

• development of explainable AI to better understand system decisions and increase user confidence.

The Self-Learning/Active Learning approach iteratively retrains the model on the most uncertain or rare samples, gradually improving accuracy without full retraining. This allows better performance on rare classes while maintaining efficiency in large-scale HTS code classification.

Thus, further implementation of semantic and multimodal models in customs classification has the potential to become an effective tool for modernizing international trade processes and increasing the overall efficiency and transparency of customs operations.

REFERENCES

1. “Best Approaches for CUSTOMS code prediction”. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.* – Available from: https://www.esann.org/sites/default/files/proceedings/2023/ES2023-163.pdf. – [Accessed: 14 Sep 2025].

2. “Neural machine translation for harmonized system codes”. *ACM Transactions on Computational Logic.* – Available from: https://dl.acm.org/doi/fullHtml/10.1145/3468891.3468915. – [Accessed: 14 Sep 2025].

3. Megdadi, E., Mohamed, A., Shaalan, K. “Machine Learning-Driven best-worst method for predictive maintenance in industry 4.0”. *Automation.* 2025; 6 (4): 91, https://www.scopus.com/pages/publications/105025823535. DOI: https://doi.org/10.3390/automation6040091.

4. He, M. “A commodity classification framework based on machine learning.” *Symmetry*. 2021; 13 (6): 964, https://www.mdpi.com/2073-8994/13/6/964. DOI: https://doi.org/10.3390/sym13060964.

5. Ding, L. “Auto-Categorization of CUSTOMS code using background net approach”. *Procedia Computer Science.* 2015; 60: 1462–1471, https://www.scopus.com/pages/publications/84941051083. DOI: https://doi.org/10.1016/j.procs.2015.05.220.

6. Sitisara, S., Jinarat, S., Ngamsaard, W. & Suthikarnnarunai, N. “Revolutionizing Harmonized system (CUSTOMS) code search with semantic search and word embeddings”. *Journal of International Trade and Economic Development*. 2025; 34 (3): 1–12. DOI: https://doi.org/10.30564/fls.v7i10.10822.

7. “Exploring machine learning models to predict harmonized system code”. *British University in Dubai Repository*. 2020. – Available from: https://bspace.buid.ac.ae/buid_server/api/core/bitstreams/302bbdc9-54c9-4b4e-9445-5d9bbe91efbf/content. – [Accessed: Apr 2020].

8. “Zonos Classify generates harmonized (CUSTOMS) codes”. *Zonos Documentation*. – Available from: https://zonos.com/docs/supply-chain/classify/classifying-goods.

9. “Feedback on Tariffy – Assistant for CUSTOMS Code Classification”. *Reddit*. – Available from: https://www.reddit.com/r/CustomsBroker/comments/1biknak/feedback_on_tariffy_assistant_for_customs_code. – [Accessed: Nov 2024].

10. Krupa, S. & Krivenchuk, Yu. “Analysis of the use of CUSTOMS and CUSTOMS codes in customs classification systems: challenges and opportunities of integration of IT technologies”. *Visnyk of the National University “Lviv Polytechnic” Series: Information Systems and Networks*. 2024; 16: 237–250. DOI: https://doi.org/10.23939/sisn2024.16.237.

11. Chen, H., van Rijnsoever, B., Molenhuis, M., van Dijk, D., Tan, Y.-H. & Rukanova, B. “The use of machine learning to identify the correctness of CUSTOMS code for the customs import declarations”. *Delft University of Technology*. 2021. DOI: https://doi.org/10.1109/DSAA53316.2021.9564203.

12. Marra de Artiñano, I., Riottini Depetris, F. & Volpe Martincus, C. “Automatic product classification in international trade: Machine learning and large language models”. *Inter-American Development Bank.* 2023. DOI: https://doi.org/10.18235/0005012.

13. Amel, O., Stassin, S., Mahmoudi, S. A. & Siebert, X. “Multimodal approach for Harmonized system code classification”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2406.04349.

14. Li, J., Wang, H. “Automatic Classification of International Trade Products Using Deep Learning”. *Journal of Applied Artificial Intelligence*. 2023; 37 (7): 556–568. DOI: https://doi.org/10.1080/08839514.2023.2198712.

15. Mall, P. K., Kumar, M., Kumar, A., Gupta, A., Srivastava, S., Narayan, V., Chauhan, A. S. & Srivastava, A. P. “Self-Attentive CNN+BERT: An approach for analysis of sentiment on movie reviews using word embedding”. *International Journal of Intelligent Systems and Applications in Engineering*. 2024; 12 (12s): 612–623, https://www.scopus.com/pages/publications/85185299520. DOI: https://doi.org/10.1109/ACCESS.2022.3154876.

16. Novak, P. & Horak, L. “Multilingual models for customs data classification”. *Information Processing & Management.* 2023; 60 (2): 102–115. DOI: https://doi.org/10.1016/j.ipm.2022.102983.

17. Singh, R., Mehta, T. “Hierarchical loss functions for commodity classification”. *Expert Systems with Applications.* 2023; 213: 118–130. DOI: https://doi.org/10.1016/j.eswa.2022.118130.

18. Petrenko, V. “Explainable AI in Customs Automation”. *Telecommunication and Information Technologies.* 2021; 25 (4): 45–56. DOI: https://doi.org/10.3318/TIT.2021.25.4.45.

19. Zhao, S. “Self-Learning Algorithms for CUSTOMS Code Assignment”. *Computers & Industrial Engineering.* 2023; 172: 108–119. DOI: https://doi.org/10.1016/j.cie.2022.108119.

20. “ATLAS: Benchmarking and adapting LLMs for global trade via Harmonized tariff code classification”. *arXiv preprint*. 2025. DOI: https://doi.org/10.1016/j.ipm.2022.102983.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 22.12.2025
Received after revision 11.02.2026
Accepted 19.02.2026

DOI: https://doi.org/10.15276/aait.09.2026.06
УДК:004

Моделі семантичного аналізу описів товарів для автоматичного визначення митних кодів

Крупа Степан Миколайович[1)]
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
Кривенчук Юрій Павлович[1)]
ORCID: https://orcid.org/000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua
[1)] Національний університет «Львівська політехніка», вул. Ст. Бандери, 12. Львів, 79013, Україна

АНОТАЦІЯ

У сучасних умовах глобальної торгівлі точна й швидка класифікація товарів за кодами Гармонізованої системи митниці має ключове значення для ефективного функціонування митних процесів. Збільшення обсягів даних та ускладнення описів товарів створюють потребу у впровадженні інтелектуальних методів аналізу текстової інформації. Використання технологій обробки природної мови у поєднанні з машинним навчанням відкриває нові можливості для автоматизації процесу визначення кодів митних та зменшення людського фактора у митній класифікації. **Метою дослідження** є підвищення точності автоматизованого визначення кодів Гармонізованої системи шляхом застосування сучасних моделей семантичного аналізу текстових описів товарів. Для досягнення цієї мети було поставлено завдання: проаналізувати існуючі моделі обробки природної мови, дослідити їх ефективність для задач митної класифікації, сформувати оптимальну архітектуру моделі та оцінити її переваги порівняно з традиційними алгоритмами. У роботі **застосовано методи** семантичного моделювання та машинного навчання. Використано експериментальний підхід до поєднання цих моделей із класифікаційними алгоритмами, зокрема логістичною регресією, деревами рішень та нейронними мережами. Оцінювання проводилося за показниками точності, повноти та міри. **Результати дослідження** показали, що використання контекстних ембедингів, зокрема моделі, забезпечує суттєве покращення точності автоматизованої класифікації товарів порівняно з традиційними статистичними методами. Запропонована узагальнена модель, що поєднує семантичний аналіз із машинним навчанням, дозволяє підвищити рівень коректного присвоєння кодів митних на основі текстових описів, навіть у випадках неоднозначних або неповних даних. Проведене дослідження підтвердило доцільність інтеграції технологій обробки природної мови в системи митної класифікації. **Наукова новизна** полягає у розробці гібридної моделі, яка об’єднує семантичні представлення текстів і алгоритми класифікації, що підвищує точність та ефективність автоматизованого визначення кодів. Практичне значення роботи полягає у можливості впровадження запропонованого.

Ключові слова: митна класифікація; семантичний аналіз; машинне навчання; обробка текстів; автоматизація; гармонізована система

ABOUT THE AUTHORS

Stepan M. Krupa - PhD student. Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
***Research field*:** Computer science, neural networks, voice signals, system programming, specialized computer systems

Крупа Степан Михайлович - аспірант. Національний університет «Львівська політехніка». Львів, вул. Степана Бандери 12, 79013 , Україна

Yurii P. Kryvenchuk - PhD, Associate Professor, Department of Artificial Intelligence Systems; Deputy Director for Scientific and Pedagogical Work, Institute of Computer Science and Information Technologies. Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
Research field: Industry 4.0, accumulation and high-speed transmission of large data volumes, Big Data analytics, AI-driven automation of manufacturing processes, Industrial IoT (IIoT), Artificial Intelligence (AI)

Кривенчук Юрій Павлович - кандидат технічних наук, доцент, доцент кафедри Систем штучного інтелекту; заступник директора з науково-педагогічної роботи Інституту комп’ютерних наук та інформаційних технологій. Національний університет «Львівська політехніка». вул. Степана Бандери 12. Львів,79013 , Україна