BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Revolutionizing Harmonized System (HS) Code Search with Semantic Search and Word Embeddings: Empowering Trade Classifications

Supamas Sitisara [1] , ***Supakpong Jinarat*** [2] , ***Witchayut Ngamsaard*** [1] , ***Nanthi Suthikarnnarunai*** [1*]

[1] *School of Engineering, University of the Thai Chamber of Commerce, Bangkok 10400, Thailand*

[2] *College of Engineering and Technology, Dhurakij Pundit University, Bangkok 10210, Thailand*

ABSTRACT

The Harmonized System (HS) code is a crucial component of global trade. It helps classify goods correctly so that taxes and duties can be applied fairly and consistently across countries. However, many current HS code search tools rely on exact keyword matches. This often causes problems like wrong results, confusion, delays, and frustration, especially for users who don't know the exact terms to search for. These mistakes can also lead to incorrect tax charges and trade issues. This study introduces a new and innovative approach to searching for HS codes. It uses semantic search and word embedding models, advanced tools from natural language processing (NLP), to understand the meaning behind what users are asking, even if they don't use the exact right words. This approach makes the search more accurate, faster, and much easier for people to use. The study includes real examples, testing, and comparisons with traditional methods to show how this new system works better. The results clearly show that it improves both speed and accuracy, helping customs officers, brokers, traders, and regulators do their jobs more efficiently and correctly. By reducing errors and making the process smoother, this new system offers a big step forward in trade technology. It shows how artificial intelligence can help make international trade more reliable, user-friendly, and ready for the future.

Keywords: Harmonized System (HS) Code; Semantic; Word Embeddings; Natural Language Processing (NLP); Customs Broker; Machine Learning

*CORRESPONDING AUTHOR:

Nanthi Suthikarnnarunai, School of Engineering, University of the Thai Chamber of Commerce, Bangkok 10400, Thailand;
Email: nanthi_sut@utcc.ac.th

ARTICLE INFO

Received: 1 July 2025 | Revised: 9 July 2025 | Accepted: 30 July 2025 | Published Online: 25 September 2025
DOI: https://doi.org/10.30564/fls.v7i10.10822

CITATION

Sitisara, S., Jinarat, S., Ngamsaard, W., et al., 2025. Revolutionizing Harmonized System (HS) Code Search with Semantic Search and Word Embeddings: Empowering Trade Classifications. Forum for Linguistic Studies. 7(10): 356–371. DOI: https://doi.org/10.30564/fls.v7i10.10822

1. Introduction

Global trade operates as a vast and intricate ecosystem, where the accurate classification of goods is not just a procedural requirement but a critical enabler of efficiency, compliance with laws and regulations, and fair commerce. Harmonized System (HS) codes, standardized by the World Customs Organization (WCO)[1], serve as a universal language for describing products and are widely used by governments, international organizations, and the private sectors for various purposes. These include trade policy, rules of origin, monitoring of controlled goods, internal taxes, freight tariffs, international trade statistics and economic research analysis, with the collection of import duties and taxes being the primary use.

However, the vast and intricate nature of the HS code system, encompassing over 5,000 commodity groups, poses significant challenges for users. Misclassifications can lead to incorrect duty assessments, shipment delays, and compliance violations, resulting in costly ramifications for both businesses and regulatory bodies.

Understanding Current Limitations

Traditional HS code search systems depend on exact keyword matching, which struggles to handle the complexity of natural language. This often creates problems for users. A single product can be described in many ways; for example, "synthetic fabric," "man-made textile," and "polyester material" may refer to the same item, but a keyword search might not connect them. HS codes also use technical terms that differ from everyday language. For instance, someone searching for "plastic bottles" may need to know the technical term "polyethylene terephthalate containers." Additionally, language and regional differences mean that users from different backgrounds may phrase the same query differently, leading to inconsistent results.

As technology evolves, semantic search is becoming increasingly essential across various industries, including e-commerce, healthcare, and customer support, because it understands the meaning and context of words rather than just matching exact terms. In HS code classification, semantic search powered by natural language processing (NLP) and word embeddings offers a smarter, more accurate, and user-friendly solution. It reduces errors, improves efficiency, and helps make global trade smoother and more reliable, setting a new standard for modern customs and trade systems.

2. Research Objectives

1) Explore Semantic Search:

   - Investigate how semantic search techniques improve understanding of user queries beyond keyword matching.

2) Implement Word Embedding Models:

   - Utilize advanced word embedding models to capture semantic relationships between queries and HS code descriptions.

3. Literature Review

As international trade continues to grow, customs processes have become more complex and challenging. Before goods can be delivered, all import and export steps must be completed. In particular, the declaration process requires detailed information, a good understanding of the products, and compliance with the laws of both the importing and exporting countries by Quan and Khan[2].

In the past, customs authorities have depended on manual inspections and keyword-based systems to classify goods by Pawłowski[3]. These methods often caused errors, especially when product descriptions were unclear or unstructured. Because the HS code system is complex, this frequently leads to mistakes and penalties.

Traditional manual and keyword-based methods often struggle to handle the complexity of product descriptions. To standardize product classification globally, the World Customs Organization (WCO) by Clark and Bernard[4] created the Harmonized System (HS) by Arya et al.[5], a global classification system. This system employs a hierarchical structure, categorizing over 5,000 product types into chapters (2 digits), headings (4 digits), and subheadings (6 digits) by Liao et al.[6]. This helps everyone involved in international trade including customs officers, customs brokers, importers, exporters, and others, to identify products using the same standard code.

Classifying products correctly is one of the most complex parts of the customs process. Using the correct Harmo-

nized System (HS) code is crucial for ensuring trade compliance and smooth customs clearance. To help with this, machine learning systems have been developed to match product descriptions with the correct codes. However, since there are thousands of HS codes, obtaining accurate results remains challenging. Mistakes can lead to serious compliance problems and financial penalties by Harsani et al.[7].

Several studies have highlighted the role of automation in the Customs Broker Management System (CBMS) in reducing manual workload, processing delays, and human errors by Zhong[8]. For instance, Hamisi and Kileo[9] proposed an automated customs clearance system integrating document verification, Harmonized System (HS) code classification, and risk assessment. Their findings demonstrated significant reductions in processing time and error rates compared to traditional manual methods by Domingues et al.[10], Gunarathne and Kalingamudali[11] and Kosgei[12].

Recent Artificial Intelligence (AI)-driven approaches by Stassin et al.[13] utilize machine learning (ML), natural language processing (NLP), and word embeddings to automate and enhance the accuracy of HS code classification. These Artificial Intelligence (AI) models outperform rule-based systems by learning from historical trade data and dynamically adjusting to new regulatory changes by Pawłowski[3], Merkulov et al.[14] and Fedotova[15].

Recent advancements in semantic search and natural language processing (NLP) by Chen et al.[16], Yereshko et al.[17] and Orłowska and Chackiewicz[18] have introduced innovative approaches to enhance HS code classification by Bleikher et al.[19], Stassin et al.[13], Novith[20] and Yuan[21].

To solve these challenges, we developed a new model that automatically matches product descriptions to HS codes without human input. It compares the text of a product description with HS code descriptions using Doc2Vec, which turns text into numerical vectors. By measuring the similarity between the two texts, the model verifies the accuracy of the HS code and flags any errors. This helps detect fraud, improve accuracy, and supports better compliance in trade operations by Chen et al.[16] and Kavoya[22].

Information Retrieval (IR) is a computer science field that focuses on extracting useful information from large datasets. It is used to help automatically classify products under the Harmonized System (HS) code by Harsani et al.[7].

Traditional keyword-based searches often give inaccurate results because product descriptions can be unclear or inconsistent. Semantic Search and Word Embedding models enable the system to understand contextual meanings, improving Harmonized System (HS) code matching by analyzing linguistic patterns rather than exact keywords by Yuan[21].

Research by Zuccon et al.[23] demonstrates the effectiveness of word embeddings in extracting meaning from text. Their application in customs environments significantly improves precision and reduces manual search time by Hambarde & Proenca[24].

Explored the use of pre-trained Semantic Textual Similarity (STS) models through deep transfer learning for HS code classification by Stein et al.[25] and Raunak[26]. Their methodology involved extracting relevant commodity information from trade documents and applying sentence-level embeddings to match descriptions with the corresponding HS codes, thereby showcasing the potential of transfer learning in this domain. Similarly, Spichakova and Haav[27] introduced a hybrid approach that integrated textual descriptions of products with the taxonomy of HS codes. By utilizing machine learning and semantic similarity, their approach enables the detection of incorrect classifications and reduces fraud in international trade by Asudani et al.[28].

Du et al.[29] introduced HScodeNet, a neural network that improves HS code classification by analyzing both detailed and overall parts of product descriptions. In another study from Liao et al.[6] developed a model that combines ERNIE (Enhanced Representation through Knowledge Integration) with a Bidirectional Long Short-Term Memory (BiLSTM) network and attention mechanisms to capture similar features. Both approaches significantly enhance classification accuracy by improving the understanding of product descriptions more effectively.

Arya et al.[5] developed a combined approach using BERT-transformer models, Named Entity Recognition (NER), distance-based methods, and knowledge graphs to classify text descriptions based on the HS code system. This comprehensive model addressed scalability and coverage challenges in HS code classification. Additionally, Chen et al.[16] approached HS code classification as a machine translation problem, modeling the translation from item descriptions to HS codes. Their method employed neural machine translation techniques, achieving substantial improvements

in classification accuracy.

Lee et al.[30] created a model using KoELECTRA for HS code classification and reached 95.5% accuracy in the top three results across 265 subheadings. These studies demonstrate that combining semantic search with machine learning, and deep learning can significantly enhance HS Code accuracy and facilitate smoother international trade.

Traditional keyword searches often give inaccurate results because product descriptions can be unclear or inconsistent. To address this, the Customs Broker Management System (CBMS) utilizes semantic search and word embedding models, including Word2Vec, GloVe, and BERT. These tools help the system understand word meanings in context, making HS code matching more accurate by Zhang & Khan[31]. Research by Worth[32], Edwards et al.[33] and Wang et al.[34] demonstrate the effectiveness of word embeddings in extracting meaning from text. Their application in customs environments significantly improves precision and reduces manual search time.

4. Research Methodology

This research presents a new system that utilizes semantic search and word embeddings to enhance user access to the correct Harmonized System (HS) code. Instead of requiring exact keywords, the system understands the meaning behind the user's input, allowing people to search using natural language. This makes the search more accurate and user-friendly, even for those without technical knowledge of HS codes.

4.1. Traditional Keyword-Based Search

Traditional HS code search tools rely on matching exact keywords entered by users with words in a database. For example, if someone searches for "plastic bottles," the system looks for those same words and ranks results based on how often they appear. While this method is simple, fast, and cheap to implement, it has many limitations.

Limitations of Keyword Search

Keyword searches are only effective when users know the exact terms used in the database. They often fail when:

- Synonyms or alternative terms are used. For instance, "synthetic leather" might not be matched with "artificial leather."
- Context is missing, leading to confusion between similar terms (e.g., "seal" as an animal vs. a mechanical part).
- Spelling mistakes are made, such as typing "lether" instead of "leather.
- Ambiguous or broad queries return too many irrelevant results.
- New ter"ms or trends are not updated in the system.
- Multilingual users input search terms in different languages, which the system cannot process.

As a result, traditional search methods can be slow, inaccurate, and frustrating, especially with large datasets or unclear product descriptions. To overcome these issues, the proposed system utilizes AI-powered semantic search to comprehend meaning, rather than merely matching words, resulting in improved performance and user satisfaction.

Semantic Search – Smarter Query Interpretation

Users can type product descriptions in everyday language, and the system intelligently interprets them to suggest the correct HS codes. This approach reduces reliance on specialized knowledge and improves accessibility across different user groups.

Word Embeddings – Context-Aware Classification

The engine understands the context and meaning behind product descriptions, ensuring precise HS code assignment. By recognizing synonyms, spelling variations, and related terms, it delivers consistent and reliable classification results (**Figure 1**).

Traditional keyword-based search systems face inherent limitations in their ability to recognize variations in terminology. For instance, shown **Figure 2**, if a user searches for the HS code for "Hom Mali Rice" or "Jasmine Rice," the system's performance hinges on the exact keywords entered. In such a setup:

Cr. Customs Department, Thailand.

- When the term "Hom Mali Rice" is used, the system may successfully identify the correct HS code because the exact keyword exists within its database.
- However, in **Figure 3**, if the user searches for "Jasmine Rice," the system may fail to recognize the term as

synonymous with "Hom Mali Rice." This limitation can result in incomplete or inaccurate results, affecting tariff classification and potentially leading to incorrect declarations.

Cr. Customs Department, Thailand.

Such challenges highlight the shortcomings of traditional keyword search, where the lack of semantic understanding can hinder the accuracy and reliability of the classification process. This underscores the need for advanced solutions, such as semantic search powered by word embedding models, to bridge the gap and ensure precise results regardless of variations in terminology.

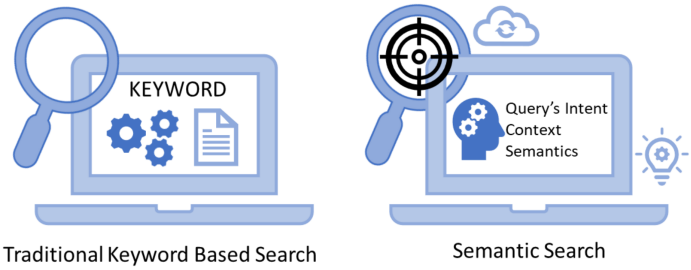

Figure 1. From Keyword Matching to Meaning: Advancing Search with AI-Powered Semantics.

Figure 2. With Traditional keyword-based Search require exact keyword in searching like "Hom Mali Rice".

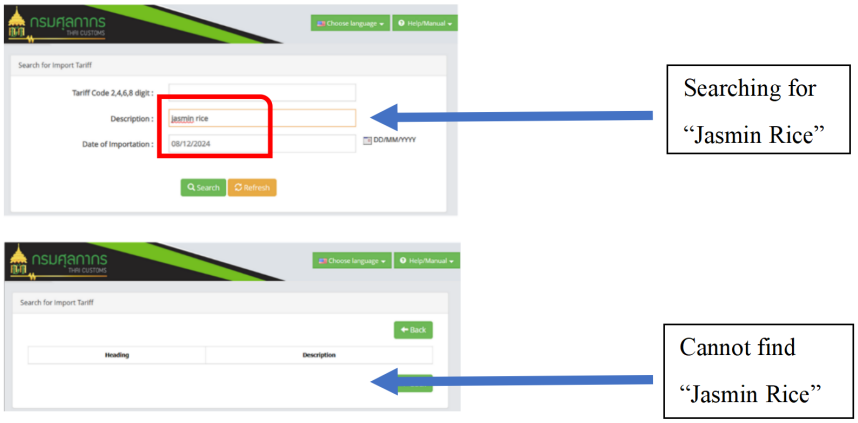

Figure 3. At traditional keyword searches, the system cannot recognize words that represent the same meaning, such as "Jasmin Rice" and "Hom Mali Rice." It can only recognize the exact word.

4.2. Exploring Semantic Search

Semantic search is an advanced method that enables systems to understand the meaning and context of words, rather than simply matching exact keywords. This is especially useful for complex tasks like HS code classification, where product descriptions can vary. Unlike traditional searches, semantic search looks at the intent behind a query, making results more accurate and relevant. This method uses Natural Language Processing (NLP) and word embedding models, which turn words into numbers (vectors) that reflect their meaning. For example, in this system, the words "cotton" and "fabric" are placed close together because they are related in meaning. This allows the system to better understand what users are really asking, even if they use different words.

4.2.1. Why Semantic Search Matters

Semantic search helps solve common problems in traditional searches:

1. Synonyms and Ambiguity—It recognizes that words like "car" and "automobile" mean the same thing, or that "bank" could mean a financial institution or a riverbank, depending on context.
2. User Differences—People may describe the same product differently. One might say "polyester fabric", while another says "synthetic textile."
3. Technical Language—HS codes often use complex terms, like calling a "plastic cup" a "molded polymer container." Semantic search helps bridge that gap.

4.2.2. How It Works

1. Natural Language Processing (NLP): This helps the system break down and understand queries. It:
    - Splits the query into words (tokenization),
    - Reduces words to their base form (like "running" → "run"), and
    - Understands how the words relate to each other.
2. Semantic Parsing: This helps the system understand the whole meaning of a sentence, not just individual words. For example, if someone searches for "HS code for exporting ceramic tableware," the system understands the topic and returns results like "earthenware dining articles," even if those exact words weren't used.

4.3. Implement Word Embedding Models

Word embeddings are a method for converting words into numerical representations, enabling computers to comprehend their meanings and relationships. These numbers, called vectors, are placed in a multi-dimensional space where similar words are located close to each other. This helps computers recognize when different words mean similar things, which is essential for accurate HS code classification.

4.3.1. Understanding Word Similarity

In word embedding models, each word is turned into a vector with many dimensions, often hundreds or more. These vectors are trained on large sets of text, allowing the system to learn which words commonly appear together. For example, words like "cotton" and "fabric" are often used in similar contexts, so their vectors end up close together. Meanwhile, "cotton" and "engine" would be far apart because they aren't related.

This enables the system to comprehend meaning, even when different words are used. So, if someone searches for "synthetic textile," the system can still find HS codes linked to terms like "polyester fabric" or "artificial cloth."

4.3.2. Capturing Language Details

Each dimension in the word vector captures a different aspect of meaning, such as material type, usage, or synonyms. For instance, the word "leather" might link to terms like "fabric" or "upholstery" depending on the context. Similarly, "synthetic leather" and "faux leather" are recognized as having similar meanings. By using these models, the system can better understand the user's intent and return more accurate HS code results, even if the wording isn't exact.

4.3.3. Implication for HS Code Search

This dimensional analysis enhances the system's ability to handle synonyms, polysemy, and domain-specific jargon. For example, a user searching for "artificial fabric" might be led to the HS code for "synthetic textiles" because the embedding model recognizes their semantic equivalence.

4.4. System Development and Tools Used

The system was implemented using Python 3.10, with the support of widely used open-source NLP libraries:

- spaCy—for text preprocessing (tokenization, lemmatization, and stopword removal),
- Gensim—for implementing the Doc2Vec model to learn vector representations of HS code descriptions,
- scikit-learn—for calculating cosine similarity and evaluating ranking performance metrics.

All experiments were performed in Google Colab Pro.

1. Dataset and Vector Construction
   Two separate datasets were used during the study:
2. Doc2Vec Training Dataset
   We utilized a publicly available dataset titled Trade Classification Data from Harvard University from Kaggle[35] to train the Doc2Vec model. This dataset includes product descriptions and their corresponding HS codes, which helped create a rich semantic representation space.
3. Evaluation Dataset
   For experimental evaluation, we constructed a curated dataset consisting of 100 real-world product descriptions used in the internal operations of ThaiSomdej. Each product in the set contains two fields: product description and HS Code. This dataset was created specifically for this research and is publicly available at ThaiSomdej[36].

4.5. Processing Pipeline

The system's workflow includes the following stages:

1. Query Input—Users input product descriptions in free-text format.
2. Preprocessing—Standard NLP operations are applied to clean and normalize the input.
3. Query Embedding—The query is converted into a vector using a pretrained Doc2Vec model.
4. HS Code Vectorization—All product descriptions from the dataset are embedded into the same vector space.
5. Similarity Computation—Cosine similarity is computed between the query and all HS code vectors.
6. Ranking and Display—The top-k most similar HS codes are retrieved and displayed to the user.

5. Evaluation Metrics and Results

The system's performance was evaluated using Precision@k for k = 5, 10, 15, 20, and 25. The **Table 1** below summarizes the results comparing the Doc2Vec-based semantic search with a traditional keyword-based method.

Table 1. Precision@k Comparison Between Semantic Search and Keyword-Based Search.

Method	Precision@5	Precision@10	Precision@15	Precision@20	Precision@25
Semantic based (doc2vec)	0.15	0.19	0.22	0.25	0.27
Keyword based (traditional)	0.04	0.06	0.06	0.06	0.06

The semantic search method demonstrates significant improvements over the keyword-based approach in all precision levels.

Application Example:

If a user searches for "HS code for leather shoes," a BERT-based system understands the meaning and can match it to the correct code, even if the official term is "footwear with leather uppers." Word embedding models are the core of modern semantic search. They turn words into numbers that capture meaning and relationships. This helps the system understand context, recognize synonyms, and give more accurate results. As a result, HS code searches become faster, more precise, and easier for users to perform.

How Semantic Search Works:

Semantic search works by turning the user's query into a vector that captures its meaning. At the same time, all HS code descriptions are also converted into vectors. The system then compares these vectors using a method such as cosine similarity to find the closest matches, not just based on keywords, but also on meaning.

Step-by-Step: HS Code Classification Using Semantic and Word Embedding Search:

The following presents the operational details corresponding to the steps illustrated in **Figure 4**.

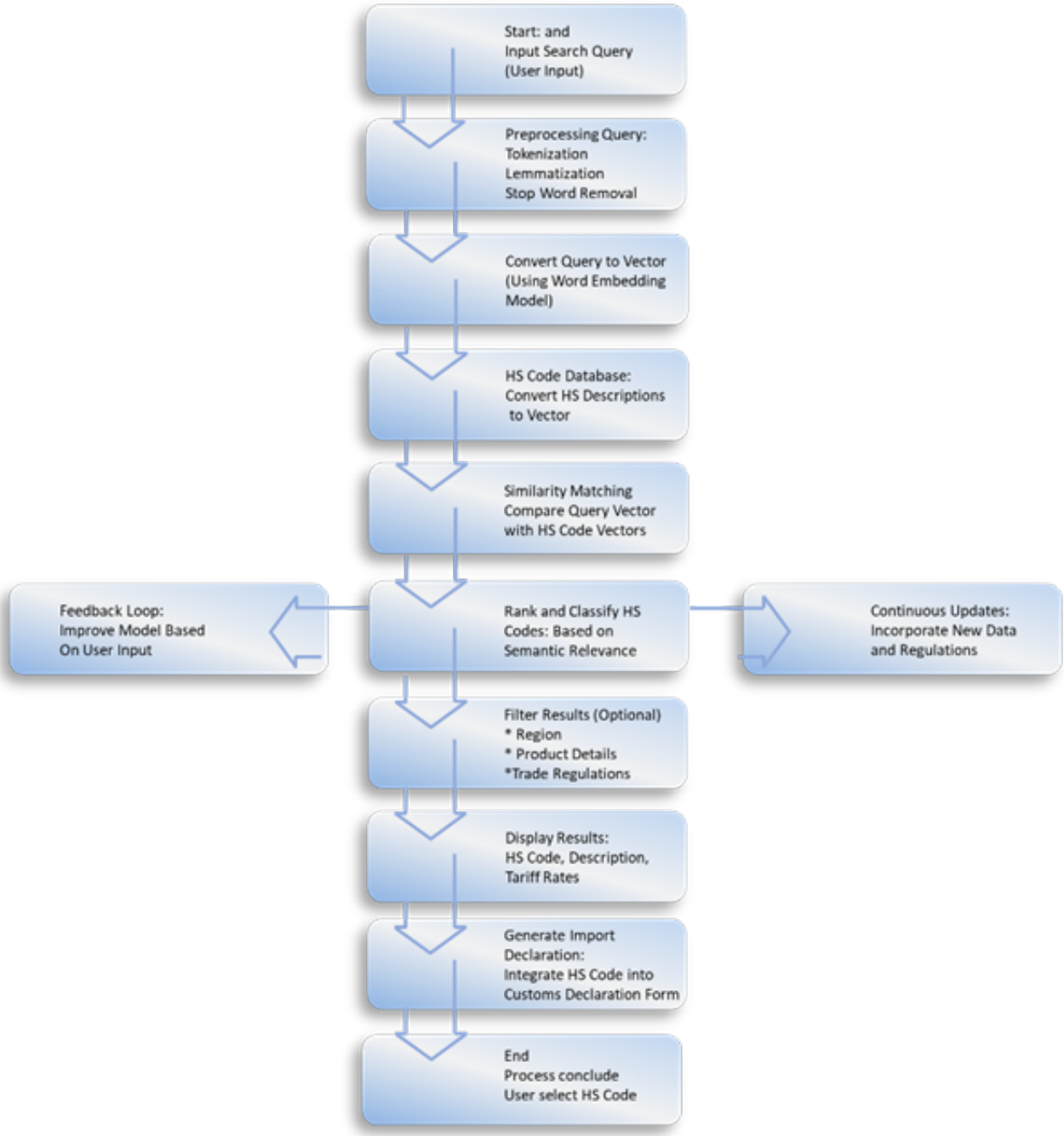

Figure 4. Full Process for HS Code Classification Using Semantic Search and Word Embeddings.

1. **Start**
   The system is activated, and the user begins by entering a product-related query.
2. **User Input**
   The user types a question in natural language, like "HS code for synthetic leather bags" or "Jasmine rice import code."
3. **Preprocessing**
   The system prepares the query by:
   - Splitting it into words (tokenization),
   - Reducing words to their base form (e.g., "running" → "run"),
   - Removing unnecessary words like "the" or "of."
4. **Convert Query to Vector**
   The cleaned query is turned into a numeric vector using

a word embedding model like Word2Vec, GloVe, or BERT. This captures the meaning of the words.

5. **HS Code Data Preparation**
   All HS code descriptions are also preprocessed and converted into vectors using the same model.
6. **Matching the Vectors**
   The system compares the query vector with the HS code vectors using similarity calculations (like cosine similarity) and finds the best matches.
7. **Ranking Results**
   The system refines the list based on region, trade context (import/export), and product details.
8. **Showing Results**
   The top matching HS codes are shown, including:
   - Code number,
   - Description,
   - Related info like tariffs and regulations.
9. **User Feedback (Optional)**
   Users can give feedback, such as "Correct" or suggest a better match. This helps improve the system.
10. **Ongoing Updates**
    The system is regularly updated to include new codes, regulations, and improve accuracy based on user input.
11. **End**
    The user selects the right HS code or adjusts the search if needed.

With semantic search, the system goes beyond traditional exact keyword matching. Instead of requiring the user to type the precise wording of a tariff description, the search engine understands the meaning and context of the query. The system then displays these possible matches, often with relevance scores (shown by the progress bars on the right).

Example:

If someone searches for "plastic beverage container," the system may return HS codes for "polyethylene bottles" or "PET containers" because it understands they are related.

For HS codes, a search for "Hom Mali Rice" doesn't have to be exact. The system knows that "Jasmine Rice" is a related term and gives the correct code shown in **Figures 5** and **6**. This saves time and helps users get accurate results without needing technical terms.

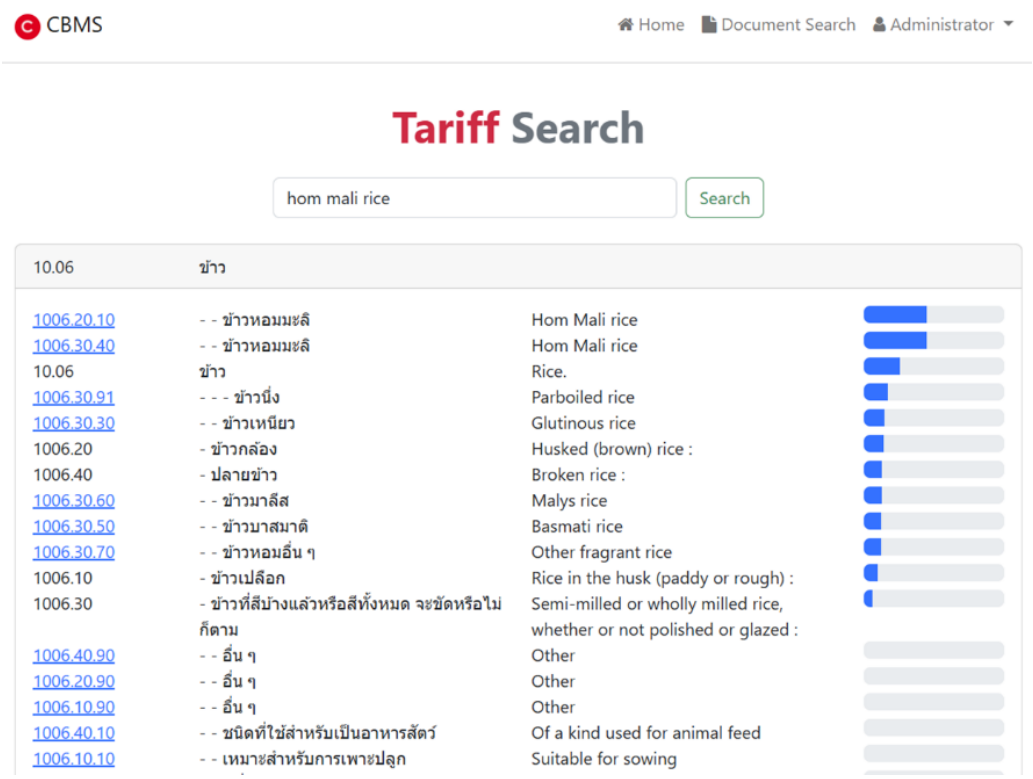

Figure 5. Searching with "Hom Mali Rice" keyword (Exact Keyword). With blue-bar, indicate the keyword search matching rate.

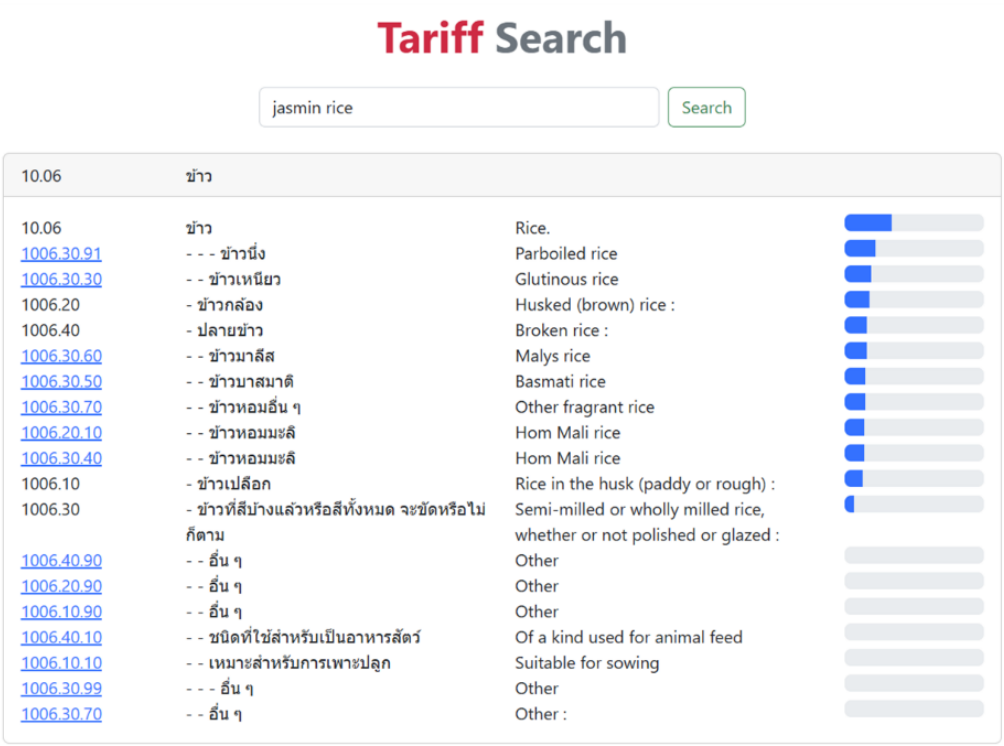

Figure 6. Searching with "Jasmine Rice" (Synonym) keyword.

With Semantic search and Word Embedding, the system does not rely solely on exact word matches. Instead, it understands the meaning and context of the query. This means that even if users enter slightly different terms, misspellings, or synonyms, the system can still identify the correct tariff codes.

Semantic search, powered by word embeddings, changes how HS codes are found by focusing on meaning, context, and synonyms, not just exact keywords. This allows the system to better understand what users are asking, making it more accurate and easier to use. By connecting everyday language with technical HS code terms, this approach improves efficiency, reduces mistakes, and makes the process smoother for businesses handling complex classifications.

Comparison between Semantic Search and Traditional Keyword Search:

Semantic search offers big advantages over traditional keyword search. It's more accurate, understands context better, and works more efficiently, making it ideal for HS code classification. However, it does require more data, advanced technology, and expert setup. Keyword search is easier and cheaper to start with, but it struggles with complex queries and isn't as reliable for today's fast-changing needs.

As illustrated in **Table 2**, several critical criteria play a pivotal role in enhancing the effectiveness of a search system. These criteria, such as time efficiency, cost-effectiveness, accuracy, relevance, and the ability to identify similar product details, are essential for achieving successful search outcomes.

Table 2. Fundamental differences between semantic search and traditional keyword-based search in tariff classification.

Factor	Semantic Search	Traditional Keyword Search
Factor	Interprets and understands the user's intent	Matches exact keywords in the query
Context Awareness	Considers the context of the query for deeper understanding	Ignores context, relying solely on word matches
Understanding Intent	Accurately recognizes the intent behind the search	Provide a literal interpretation of the query
Results Accuracy	Delivers relevant and user-centric results	May yield irrelevant or imprecise outcomes
Time	Faster retrieval due to optimized understanding of intent and context	Slower due to reliance on trial-an-error query reformulation

Table 2. *Cont.*

Factor	Semantic Search	Traditional Keyword Search
Cost	Higher initial investment due to advanced algorithms and implementation	Lower initial cost but higher maintenance cost due to inefficiencies
Barriers	Requires robust datasets and computational resources for effective deployment	Limited capability to handle complex queries or ambiguous inputs
Error Rate	Lower error rate due to contextual understanding and intent recognition	Higher error rate due to reliance on exact matches and lack of nuance
Workforce	Requires skilled personnel for initial setup, model training, and maintenance	Easier for less specialized teams to manage, but may require more manual adjustments
Adaptability	Highly adaptable to evolving terminology and user behaviors	Limited adaptability; requires frequent manual updates to maintain relevance
Learning Curve	Steeper learning curve due to the complexity of the system	Relatively straightforward to learn and use
Scalability	Easily scalable to accommodate larger datasets and diverse languages	Scalability is challenging due to inefficiencies with large datasets
Maintenance	Requires periodic retraining and updates for models	Requires regular keyword updates and rules modification
Search Precision	Provides highly precise and relevant results through semantic understanding	Results may be less precise, especially for ambiguous or complex queries
System Performance	High performance in retrieving accurate results quickly, even with complex or ambiguous queries	Performance degrades with complex queries or large datasets, leading to slower responses
Operational Efficiency	Streamlines workflows by automating complex classification processes, reducing manual intervention	Requires more manual adjustments and corrections, leading to inefficiencies in operations
User Experience	Delivers a more intuitive and user-friendly experience, enabling natural language queries and better engagement	Can frustrate users due to rigid keyword requirements and irrelevant results for unclear queries

Case Study

This independent case study involves no conflict of interest and is conducted with permission from ThaiSomdej Service Company Limited to disclose and publish its content. ThaiSomdej Service Company Limited, established in 1970, is a well-known customs brokerage firm in Thailand. The company developed this innovation to improve customs operations and enhance service quality, particularly in HS code classification. To overcome the limitations of traditional keyword-based searches, a new search model was introduced using semantic search and word embedding technology. This approach offers a smarter, more accurate, and more efficient way to classify HS codes. **Table 3** illustrates the practical differences between the two approaches. Traditional keyword search successfully identifies results only when exact terms are used, but it fails when queries involve synonyms, misspellings, or broader product descriptions. For example, while "Hom Mali Rice" is found, the synonymous term "Jasmine Rice" is not recognized. Similarly, high-dimensional terms like "Artificial Fabric" and brief or variant keywords such as "Vaccines for animals" are overlooked. By contrast, the CBMS with semantic and word embedding search consistently delivers accurate results across all cases, including synonyms, complex terms, and spelling errors.

This case study clearly demonstrates that the CBMS significantly enhances accuracy, accessibility, and reliability in HS code classification, transforming what was once a rigid, error-prone process into a smarter, more user-friendly solution.

Key Observations:

- **Exact Matches:** Both systems found results for terms like "Hom Mali Rice" and "Benzyl Alcohol." However, the traditional search failed on "Plastic Bottles," while CBMS succeeded.
- **Similar Meaning:** The traditional system missed related terms like "Jasmine Rice" (same as "Hom Mali"),

but CBMS recognized the connection.

- **Contextual Terms:** For complex terms like "Artificial Fabric," only CBMS gave results. Both systems worked for "Synthetic Textile."
- **Misspellings:** The traditional search didn't recognize errors like "Dioxine" (intended: "Dioxide"), but CBMS corrected it.
- **Short or Synonym Terms:** CBMS handled short terms like "Vaccines for vet" and "Self-adhesive film," while the traditional search failed. It also understood synonyms like "Vaccines for animals."

The comparison highlights a significant performance gap between traditional keyword search and CBMS enhanced with semantic and word embedding search.

- Traditional Keyword Search requires exact keyword matches, which often leads to failed results if users misspell terms, use synonyms, or enter incomplete/complex queries. This limitation is reflected in its 40% success rate (4 out of 10 queries), showing that more than half of the searches did not return correct or useful results.
- CBMS with Semantic and Word Embedding Search**,** on the other hand, interprets the meaning and context of search terms rather than relying on exact matches.

For the results in **Table 4**, CBMS with Semantic and Word Embedding Search achieved a 100% success rate, far better than the 40% from traditional keyword search. Its key strengths include understanding synonyms, fixing misspellings, and handling short or complex keywords.

This advanced search capability enhances accuracy, speed, and efficiency in identifying tariff classifications, delivering superior performance and user satisfaction.

The contrasts Traditional Keyword Search with Semantic and Word Embedding Search (as used in CBMS) across 10 performance criteria:

1. Accuracy: Both methods can provide accurate results, but semantic search ensures accuracy more consistently by interpreting meaning.
2. Context Awareness: Traditional search cannot understand context, while semantic models capture relationships and meanings between words.
3. Synonyms & Similar Words: Traditional systems fail if the exact term is not used; semantic models recognize synonyms (e.g., "car" vs. "automobile").
4. Misspellings: Traditional search rejects misspelled queries; semantic models can still interpret and return relevant results.
5. Complex Keywords: High-dimensional or multi-word queries confuse traditional search, but semantic models handle them effectively.
6. Ease of Finding Results: Users often struggle with traditional search; semantic search makes it easier by aligning results with intent.
7. Reference Information: Both approaches can provide reference details once results are found.
8. Speed: Traditional search can be slower when queries don't match exactly, requiring retries; semantic search retrieves the right results faster.
9. Cost: Traditional search is cheaper to implement and maintain, whereas semantic search requires higher investment in models and infrastructure.
10. Maintenance: Traditional systems are simpler; semantic models require ongoing updates, training, and higher technical maintenance.

From the results in **Table 5**, Semantic and Word Embedding Search consistently outperforms traditional keyword search. It delivers higher accuracy, faster results, and handles synonyms, misspellings, and complex queries better. These strengths make CBMS the preferred and more effective system for HS code classification.

Table 3. The comparison table highlights the performance of two search systems: the Traditional Keyword Search and the Semantic with Word Embedding Search.

No.	Sample Keyword Search	Word Type	Customs Department Tariff Classification with Traditional Keyword Search	Customs Broker Management System (CBMS) with Semantic and Word Embedding Search
1	Horn Mali Rice	Exact Keyword	Found	Found
2	Jasmine Rice	Similar Meaning Keyword (Jasmine = Horn Mali)	Not Found	Found

Table 3. *Cont.*

No.	Sample Keyword Search	Word Type	Customs Department Tariff Classification with Traditional Keyword Search	Customs Broker Management System (CBMS) with Semantic and Word Embedding Search
3	Benzyl Alcohol	Exact Keyword	Found	Found
4	Plastic Bottles	Exact Keyword	Not Found	Found
5	Artificial Fabric	High-Dimensional Keyword	Not Found	Found
6	Synthetic Textile	High-Dimensional Keyword	Found	Found
7	Pigments with titanium Dioxine not more than 80%	Misspelling Keyword (correct spelling = Dioxide)	Not Found	Found
8	Vaccines for vet	Brief Keyword	Found	Found
9	Vaccines for animals	Synonym Keyword	Not Found	Found
10	Self-adhesive film	Brief Keyword	Not Found	Found

Table 4. The comparison highlights a significant performance gap between traditional keyword search and CBMS enhanced with semantic and word embedding search.

Search Method	Successful Matches	Total Queries	Success Rate
Traditional Keyword Search	4	10	40%
CBMS with Semantic and Word Embedding Search	10	10	100%

Table 5. Comparing between Normal Search and Searching with Semantic and Word Embedding.

No.	Statement	Traditional Search	Semantic and Word Embedding Search
1	Accurate Search Results	Yes	Yes
2	Understanding Context	No	Yes
3	Understand Synonyms or Similar words	No	Yes
4	Recognize Misspelling	No	Yes
5	Understand High-Dimensional or Complex Keywords	No	Yes
6	Find Search Result Easy	No	Yes
7	Provide References Info	Yes	Yes
8	Find Search Result Fast	No	Yes
9	Cheaper Cost	Yes	No
10	High Maintenance	No	Yes

6. Limitations of Semantic Search and Word Embedding Models in HS Code Classification

Semantic search and word embedding models offer major improvements over keyword-based systems, but they still have some challenges. Understanding these limitations helps ensure better use and results.

1. **Need for Quality Data**

These models depend on good training data. If the data is old or incomplete, the results may be inaccurate—especially if new HS codes like "biodegradable plastics" aren't included.

2. **Struggles with Rare Terms**

Uncommon or technical terms (like "photovoltaic cells") may not appear often in training data, so the system may not classify them correctly.

3. **Overgeneralization**

The system might return broad results. For example, searching "leather handbags" could bring up all types of leather goods, requiring users to narrow it down.

4. **High Resource Requirements**
   Advanced models like BERT need strong computing power and regular updates, which can be costly for smaller organizations.
5. **Multilingual Challenges**
   Models trained mainly in English may not work well with other languages. A French query like "riz parfumé" might not match with "Jasmine Rice."
6. **Handling Ambiguous Queries**
   Some queries have multiple meanings. For example, "plastic packaging" might match both hard containers and plastic film, causing confusion.
7. **Ongoing Maintenance**
   The system needs regular updates to stay current with new products and regulations. Without this, it may produce outdated or incorrect results.
8. **Lack of Transparency**
   These models often work like "black boxes," making it hard to explain how they reach a decision, an issue for compliance-focused users like customs authorities.

In short, while these technologies are powerful and improve HS code classification, they require careful planning, quality data, and regular updates to work well and avoid errors.

7. Area for Future Exploration and Improvement

While semantic search and word embeddings have improved HS code classification, there's still room to grow. Key areas for future development include:

1. **Better Handling of Rare Terms**
   These models sometimes struggle with uncommon or very specific terms. Training on more industry-specific data can help improve accuracy for niche products.
2. **Stronger Multilingual Support**
   Current models don't always work well with non-English queries. Future systems should be trained on multiple languages to ensure global accuracy.
3. **More Transparency**
   Deep learning models are often hard to understand. Using explainable AI (XAI) can help users see how the system makes decisions—important for customs and trade compliance.
4. **Reducing Confusion from Ambiguity**
   Semantic models can sometimes give broad or unclear results. Combining them with rule-based systems could help provide more accurate answers.
5. **Improving Efficiency**
   Powerful models like BERT use a lot of computing power. Lighter, more efficient models would make these tools more accessible to smaller organizations.
6. **Learning from User Feedback**
   Many systems don't adapt to user corrections. Future improvements could include learning from real-time feedback to improve results over time.
7. **Keeping Up with Regulation Changes**
   HS codes and trade rules often change. Automating updates to the system will help ensure the model stays current and accurate.
8. **Using Visual Data**
   Some products are hard to describe in words. Adding image recognition and visual data could improve classification—especially for items like machinery or textiles.

8. Conclusions

This new approach has a significant impact on customs authorities, businesses, and traders. Accurate HS code classification helps speed up customs procedures, reduce mistakes, and follow international trade rules more easily. It's especially helpful for small and medium-sized businesses (SMEs), which may not be familiar with complex trade terms, giving them a fairer chance in global markets.

This study tackles the main problems with traditional keyword-based searches by introducing a smarter method using semantic search and word embeddings. This innovation enhances HS code searches, making them more accurate, faster, and easier to use, thereby setting a new benchmark in customs classification systems.

In short, semantic search changes how we find HS codes by understanding the real meaning of what users are asking. By using advanced language tools like natural language processing (NLP) and word embeddings, it solves the weaknesses of old systems and offers a better, more reliable way to classify products in global trade.

Author Contributions

Conceptualization, S.S. and N.S.; software, S.J.; validation, S.S., W.N. and N.S.; data curation, S.J. and N.S.; formal analysis, S.S. and S.J.; writing—original draft preparation, S.S.; writing—review and editing, N.S., W.N. and S.J.; visualization, S.S.; supervision and project administration, N.S. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Written informed consent to publish this paper has been obtained from the patient(s), where applicable.

Data Availability Statement

Unavailable data due to privacy restrictions.

Acknowledgments

The authors gratefully acknowledge the administrative assistance, technical support, and provision of materials that contributed to the completion of this work.

Conflicts of Interest

The authors declare no conflicts of interest. The funders had no role in the study design, data collection and analysis, manuscript preparation, or publication decision.

References

[1] Allende, J., 2022. World Customs Organization. Springer: Cham, Switzerland.

[2] Quan, J., Khan, M.S., 2024. The mediating role of job satisfaction and competitive advantage between quality management practices and sustainable performance: Case of hospitals in Guangxi, China. Human Systems Management. 43(6), 971–988. DOI: https://doi.org/10.3233/HSM-240045

[3] Pawłowski, M., 2022. Machine learning based product classification for ecommerce. Journal of Computer Information Systems. 62(4), 730–739.

[4] Clark, J., Bernard, D., 2022. Customs in a world of enhanced trade facilitation. In: Customs Matters: Strengthening Customs Administration in a Changing World. International Monetary Fund: Washington, DC, USA.

[5] Arya, A., Roy, S., Jonnala, S., 2023. An Ensemble-based approach for assigning text to correct Harmonized system code. In Proceedings of the 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). DOI: https://doi.org/10.48550/arXiv.2211.04313

[6] Liao, M., Huang, L., Zhang, J., et al., 2024. Enhanced HS Code Classification for Import and Export Goods via Multiscale Attention and ERNIE-BiLSTM. Applied Sciences. 14(22), 10267.

[7] Harsani, P., Suhendra, A., Wulandari, L., et al., 2020. Artificial intelligence-based methods for harmonized system code translation: A review. Journal of Advanced Research in Dynamical and Control Systems. 12(2), 1389–1398.

[8] Zhong, C., 2024. AI-Powered Customs Clearance: Optimizing Trade Compliance and Border Management. Journal of AI-Driven Trade Facilitation Engineering and Single Window Systems. 2(1), 79–98.

[9] Hamisi, S.R., Kileo, W.J., 2024. The Effect of Automated Customs Clearance Systems on Enhancing Trade Efficiency in Tanzania. International Journal of Social Sciences and Management Research. 10(8), 408–422. DOI: https://doi.org/10.56201/ijssmr.v10.no8.2024.pg408.422

[10] Domingues, P., Carreira, P., Vieira, R., et al., 2016. Building automation systems: Concepts and technology review. Computer Standards & Interfaces. 45, 1–12.

[11] Gunarathne, S., Kalingamudali, S., 2019. Smart automation system for controlling various appliances using a mobile device. In Proceedings of the 2019 IEEE International Conference on Industrial Technology (ICIT), Melbourne, VIC, Australia, 13–15 February 2019.

[12] Kosgei, S.K., 2019. Effect of automated customs procedures on trade facilitation a case of clearing and forwarding agents in Nairobi region. KESRA/JKUAT: Juja, Kenya.

[13] Stassin, S., Amel, O., Mahmoudi, S., et al., 2023. Similarity versus Supervision: Best Approaches for HS Code Prediction. ESANN. 175–180.

[14] Merkulov, R., Chien, V., Khodaverdian, A.E., et al., 2023. Machine learning based product classification and approval. USA. 20230252544. 10 August 2023.

[15] Fedotova, G., 2020. Problems of digital transformation of customs services on classification of goods. In Proceedings of the 2nd International Scientific Conference on Innovations in Digital Economy; pp. 1–10. DOI: https://doi.org/10.1145/3444465.3444503

[16] Chen, X., Bromuri, S., Van Eekelen, M., 2021. Neural machine translation for harmonized system codes prediction. In Proceedings of the 2021 6th International Conference on Machine Learning Technologies; pp. 158–163. DOI: https://doi.org/10.1145/3468891.3468915

[17] Yereshko, K., Khoma, O., Pyslytsia, A., 2024. Digitalization of Customs Procedures: Current State and Prospects. Journal of Vasyl Stefanyk Precarpathian National University. 11(2), 103–115.

[18] Orłowska, M., Chackiewicz, M., 2024. Logistics and Customs Handling–New Technologies and Operational Efficiency and Compliance with International Regulations. Scientific Papers of Silesian University of Technology. Organization & Management. (211), 499–514.

[19] Bleikher, O.V., Ageeva, V.V., Brazovskaya, O.E., et al., 2016. Using information logistics techniques to develop an integrated information pool for improving efficiency of post-clearance customs control. Information Technologies in Science, Management, Social Sphere and Medicine. DOI: https://doi.org/10.2991/itsmssm-16.2016.32

[20] Novith, D.C., 2024. Harmonized System Code Recommendation: A Multi-Class Classification Model. Jurnal BPPK: Badan Pendidikan dan Pelatihan Keuangan. 17(3), 1–11.

[21] Yuan, Y., 2020. Improving information retrieval by semantic embedding [Master Thesis]. University of Twente: Enschede, Netherlands.

[22] Kavoya, J., 2020. Machine learning for intelligence driven Customs management. African Tax and Customs Review. 1(3), 50–58.

[23] Zuccon, G., Koopman, B., Bruza, P., et al., 2015. Integrating and evaluating neural word embeddings in information retrieval. In Proceedings of the 20th Australasian Document Computing Symposium; pp. 1–8. DOI: https://doi.org/10.1145/2838931.2838936

[24] Hambarde, K.A., Proenca, H., 2023. Information retrieval: recent advances and beyond. IEEE Access. 11, 76581–76604.

[25] Stein, R.A., Jaques, P.A., Valiati, J.F., 2019. An analysis of hierarchical text classification using word embeddings. Information Sciences. 471, 216–232.

[26] Raunak, V., 2017. Simple and effective dimensionality reduction for word embeddings. arXiv preprint. arXiv:1708.03629.

[27] Spichakova, M., Haav, H.-M., 2020. Application of Machine Learning for Assessment of HS Code Correctness. Baltic Journal of Modern Computing. 8(4), 698–718.

[28] Asudani, D.S., Nagwani, N.K., Singh, P., 2023. Impact of word embedding models on text analytics in deep learning environment: a review. Artificial Intelligence Review. 56(9), 10345–10425.

[29] Du, S., Wu, Z., Wan, H., et al., 2021. HScodeNet: Combining hierarchical sequential and global spatial information of text for commodity HS code classification. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining; pp. 676–689. DOI: https://doi.org/10.1007/978-3-030-75765-6_54

[30] Lee, E., Kim, S., Kim, S., et al., 2021. Classification of goods using text descriptions with sentences retrieval. arXiv preprint. arXiv:2111.01663.

[31] Zhang, H., Khan, M.S., 2024. Empirical Research On Ethical Leadership And Knowledge Workers' Innovative Behaviour: The Mediating Role Of Job Autonomy. Revista de Gestao Social e Ambiental. 18(9). DOI: https://doi.org/10.24857/rgsa.v18n9-091

[32] Worth, P.J., 2023. Word embeddings and semantic spaces in natural language processing. International Journal of Intelligence Science. 13(1), 1–21.

[33] Edwards, A., Camacho-Collados, J., De Ribaupierre, H., et al., 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, December 2020; pp. 5522–5529.

[34] Wang, C., Nulty, P., Lillis, D., 2020. A comparative study on word embeddings in deep learning for text classification. In Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval; pp. 37–46. DOI: https://doi.org/10.1145/3443279.3443304

[35] Trade Classification Data_EDA, 2023. Kaggle Notebook. Available from: https://www.kaggle.com/code/kaggleprollc/trade-classification-data-eda (cited 18 May 2024).

[36] ThaiSomdej Dataset, 2023. Project.devplanter.com. Available from: http://project.devplanter.com/dataset.zip (cited 15 April 2023).