Explainable Product Classification for Customs

EUNJI LEE, School of Computing, KAIST, Republic of Korea
SIHYEON KIM, School of Computing, KAIST, Republic of Korea
SUNDONG KIM*, AI Graduate School, GIST, Republic of Korea
SOYEON JUNG, ICT and Data Policy Bureau, Korea Customs Service, Republic of Korea
HEEJA KIM, Customs Valuation and Classification Institute, Korea Customs Service, Republic of Korea
MEEYOUNG CHA, School of Computing, KAIST, Republic of Korea

The task of assigning internationally accepted commodity codes (aka HS codes) to traded goods is a critical function of customs offices. Like court decisions made by judges, this task follows the doctrine of precedent and can be nontrivial even for experienced officers. Together with the Korea Customs Service (KCS), we propose a first-ever explainable decision supporting model that suggests the most likely subheadings (i.e., the first six digits) of the HS code. The model also provides reasoning for its suggestion in the form of a document that is interpretable by customs officers. We evaluated the model using 5,000 cases that recently received a classification request. The results showed that the top-3 suggestions made by our model had an accuracy of 93.9% when classifying 925 challenging subheadings. A user study with 32 customs experts further confirmed that our algorithmic suggestions accompanied by explainable reasonings, can substantially reduce the time and effort taken by customs officers for classification reviews.

CCS Concepts: • **Information systems → Expert systems**; • **Computing methodologies → Natural language processing**; • **Applied computing → E-government**.

Additional Key Words and Phrases: Product classification, Interpretability, Decision support, Human-centered explainable AI

1 INTRODUCTION

With the continuing advances in artificial intelligence, computational models are now being used to automate not only simple laborious tasks but also complex tasks that once seemed irreplaceable by machines. One example is self-driving cars, which are now available from a myriad of brands like Tesla and Google. AI is taking over the mundane task of steering the car, and rapidly learning to handle unknown scenarios. In legal sectors, tribunal and court decisions are being assisted by AI [31]. Many other domains are adopting AI in their core functions, including medical decisions, surveillance, climate modeling, and financial predictions.

However, it remains unclear whether AI can completely replace human tasks. In particular, some argue that AI shouldn't be a final arbiter for mission-critical tasks that require human reasoning [24]. For example, while court decisions must be based on an in-depth understanding of the precedent and relevant laws; they are also subject

*Corresponding author: sundong@gist.ac.kr

Authors' addresses: Eunji Lee, mk35471@gmail.com, School of Computing, KAIST, 291 Daehak-ro, Daejeon, Republic of Korea, 34141; Sihyeon Kim, School of Computing, KAIST, Daejeon, Republic of Korea, sihk@kaist.ac.kr; Sundong Kim, AI Graduate School, GIST, 123 Cheomdangwagi-ro, Gwangju, Republic of Korea, sundong@gist.ac.kr; Soyeon Jung, ICT and Data Policy Bureau, Korea Customs Service, Daejeon, Republic of Korea, jsy6519@korea.kr; Heeja Kim, Customs Valuation and Classification Institute, Korea Customs Service, Daejeon, Republic of Korea, tart75@korea.kr; Meeyoung Cha, School of Computing, KAIST, Daejeon, Republic of Korea, meeyoung.cha@kaist.ac.kr.

ACM 2157-6904/2023/12-ART
https://doi.org/10.1145/3635158

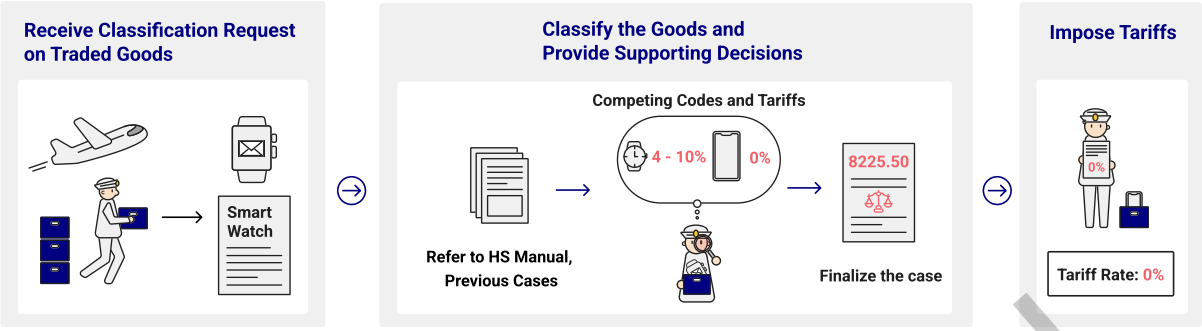


Fig. 1. Overview of the HS code classification procedure at the customs office.

to subtlety and sometimes need moral and policy judgments. Others have argued that, for this reason, outcomes of data-driven models cannot be perceived as objective by the public nor replace field experts [9].

Explainable AI (XAI) is a promising alternative to conventional AI, for use as an assistance tool in various sectors. XAI offers human interpretable reasoning with algorithmic suggestions and can assist humans in taking an arbiter role. At the same time, it can substantially reduce the time and effort needed for humans to complete complex tasks. In many sectors, when implementing AI in high-stakes scenarios, XAI is considered a critical requirement [18].

This paper presents a first-ever XAI model to assist customs officers in assigning commodity codes (aka HS codes) as they classify traded goods. As in the legal sectors, general classification guidelines are written in an HS Classification Handbook[1], which is internationally accepted. However, classification is a nontrivial task even for experienced customs officers. HS code determines the tariff borne by the importer, and hence some decisions can lead to an international dispute. Disputed cases are handled by national and international customs committees, whose decisions are later used as the doctrine of precedent. Our goal in this work is to implement an AI model that provides top suggestions along with human-interpretable reasoning.

Figure 1 shows an overview of the HS code classification procedure. Traded goods need to be declared with HS codes based on the HS manual, and the general regulations for HS. This decision is sensitive as it determines the tariff rate, yet it is not obvious, as similar goods may need to be assigned different HS codes and vice versa. Customs experts examine disputed cases, and their decisions are logged at customs offices. The primary goal of our AI model is to assist this decision process by giving algorithmic suggestions along with human interpretable reasoning.

Our AI model operates in two stages. First, it predicts item classification based on the text description of the goods. Second, it retrieves evidence about each candidate from the HS manual. As a result, the AI suggestion consists of candidate codes and the relevant key sentences in the HS manual as explainable evidence. The AI model runs on top of the latest Natural Language Processing (NLP) model and has been tested on real HS classification cases involving mechanical and electrical equipment (as listed in Chapter 84, 85, and 90), the goods categories that are known to be most challenging for human officers because of their similarity. Our model showed a high accuracy of 93.9% when the top-3 candidates of the 6-digit HS codes were suggested for 925 classes.

This work greatly benefits from our collaborating partners at the Korea Customs Service (KCS), from which we recruited 32 field officers to help test the efficacy of the prototype AI model. Customs field officers of varying career experience participated in our usability survey. The survey data indicated that the proposed AI model was perceived to be 'helpful' (by 85% of the respondents) as a supporting tool, and officers found value in its

[1]https://www.wcoomdpublications.org/en/products/harmonized-system/explanatory-notes-2022

ability to reduce the time needed for screening candidate codes. In particular, the AI model was perceived to be more helpful by officers with shorter field experience, who valued receiving multiple candidate suggestions and their interpretable evidence. These findings suggest that AI models can be used to train and assist novice customs officers and contribute to cost reduction. Currently, we are beta-testing the HS classification service for target users at Korea Customs Service: see https://ds.ibs.re.kr/product-classification/. It is worth noting that AI is increasingly providing support in various domains, such as legal cases with predictive judgment systems [34] and search systems[2]. Similarly, this work has the potential to be beneficial in the context of HS code classification tasks. In conclusion, we discuss the implications of these findings and outline potential future directions. We conclude with a discussion of implications and future directions.

Codes and implementation details can be found via the GitHub repository at: https://github.com/sprain02/HS_classification

2 PRELIMINARIES

2.1 HS Classification for Traded Products

According to the World Customs Organization (WCO), the number of import and export declarations worldwide reached 500 million in 2020. Events like the outbreak of coronavirus disease 2019 (COVID-19) have led to a surge in the cross-national imports of e-commerce goods, where for instance, Korea accounted for 63.5 million in 2020, a 48% increase compared to the previous year [29]. As global transactions increase and traded products become diversified, the management of standards for categorizing numerous products—i.e., Harmonized Commodity Description and Coding System, or Harmonized System (HS) for short— is becoming crucial. The HS is an international standard for classifying goods. From live animals to electronic devices, each product is classified under one of 5,387 subheadings (the first six digits of the HS codes) that meet international conventions [42]. This code determines critical trade decisions like tariff rates and import and export requirements.

HS code classification is nontrivial and requires a high degree of expertise since it determines the tariff rate. Securing tariffs is vital for fiscal income in many countries. The share of tax revenue secured through customs offices is nearly 20% worldwide and exceeds 40% in West African countries.[3] In addition, tariff rates are directly linked to the price of goods, affecting their global competitiveness. Therefore, both importers and exporters pay special attention to product declarations. Customs authorities scrutinize the submitted HS codes of declared goods and correct them if needed. Simple errors can be corrected by amending the declaration or sending a request for correction. If customs administrations find evidence of smuggling or deliberate false declarations for tax evasion purposes, then importers are punished by customs acts.

The process of classifying a product is complex because human interpretations may not always be consistent, which can lead to an international dispute when a ruling between customs authorities differs or differs between the companies and customs authorities. For example, when smartwatches were first released, tariffs varied across importing countries due to the absence of a classification standard. As shown in Figure 1, tariff rates for wireless communication devices are 0%, but 4–10% for watches, which led to a dispute that was finally resolved by the WCO HS Committee in 2014. The committee classified smartwatches as wireless communication devices, and the manufacturer was able to save approximately $13 million per year [33]. Furthermore, difficulties arise when the product has multiple characteristics or new characteristics that are not mentioned clearly in the guidelines. The customs administration operates a pre-examination system, allowing import and export companies to request customs to review their items before formal declarations. Korea Customs Service receives approximately 6,000 applications for pre-examination every year. With the increasing complexity of goods, the processing time has

[2]https://deepjudge.ai/

[3]WCO annual report shows the proportion of revenue collected by customs in tax revenue of each country (pp. 46-91): https://tinyurl.com/yxjvn9mz

increased from 20.4 to 25.8 days per inspection since 2018. The main reason for this is the detailed review process since the emergence of the HS code and that the corresponding tax rate can differ, even for similar-looking items. For example, tariff rates for television (HS 8528.59) are 8% but 0% for PC monitors (HS 8528.52).

The HS code classification process includes a review of the descriptions submitted by applicants and relevant cases in the past. Experts adjust to the HS manual that includes Explanatory notes of the HS [6] for standard code descriptions, General Rules of Interpretation (GRI) [40] for decision-making criteria, HS Nomenclature, and HS Compendium of Classification Opinions. We designed our model to reflect this process. First, it suggests the first six digits of HS codes (called *subheadings*) based on product descriptions with pretrained language models. Then, it retrieves key sentences from the HS manual that are most related to the product. The retrieved sentences act as supporting facts to support the final decision statement by the officer, which will be provided to the importers and exporters who request classification.

2.2 HS Code and Its Classification

The WCO explains that the Harmonized Commodity Description and Coding System (popularly known as the Harmonized System or the HS) is one of the most successful instruments ever developed by the WCO. It is a multipurpose goods nomenclature used by more than 200 countries and Customs or Economic Unions as the basis for Customs tariffs and the compilation of international trade statistics.[4]

All the items that go through customs are assigned an HS code, an internationally standardized system of names and numbers to classify traded products to determine tariffs. As an internationally recognized standard, the first six digits of the HS code (HS6) are the same for all countries. Countries have added more digits to their respective HS code systems for further classification. HS6 includes the following three components:

(1) **Chapter**: the first two digits of the HS code, which contains 96 categories from 01 to 99.
(2) **Heading**: the first four digits of the HS code, groups similar characteristics of goods within a chapter.
(3) **Subheading**: the first six digits the of HS code, groups goods within a heading.

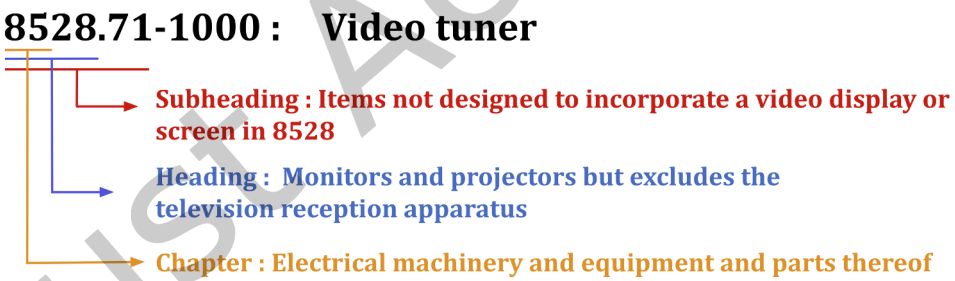


Fig. 2. Hierarchical structure of the HS code.

As a characteristic data property, we note that HS code heading and subheading data are skewed in terms of their frequency. Popular headings and subheadings appear disproportionately more times in the classification data, which could help the model learn the data traits. Figure 3 shows an example of the frequent and infrequent headings in the Chapter 85 data, where the first three headings appear over 8000 times each. The least popular item appears less than 500 times. Such popularity also has evolving temporal trends, with some headings becoming less requested (or more requested) over time. We later discuss how the skewed popularity and temporal dynamics affect quality in the Discussion section.

[4]http://www.wcoomd.org/en/topics/nomenclature/overview/what-is-the-harmonized-system.aspx

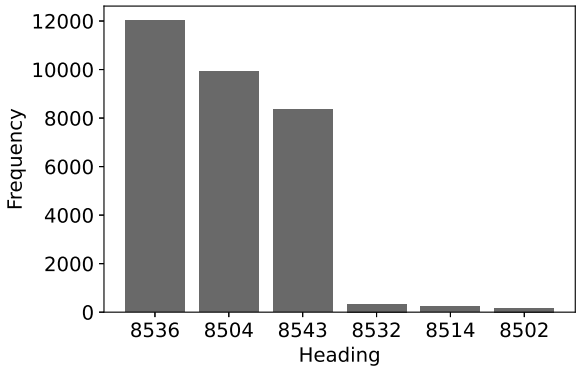


Fig. 3. Imbalanced heading data in decision case.

Recent studies have utilized machine learning approaches to classify HS codes using text descriptions of the declared goods. These approaches include the k-nearest neighbor, support vector machine (SVM), Adaboost [10], and neural networks [3]. For text embedding, BERT [14] and ELECTRA [7] have been introduced. To capture semantic information, studies have used neural machine translators [1] and other transformer-based algorithms [22]. Some have utilized hierarchical relationships between the HS codes and the co-occurrence of the words using background nets [21]. Similar studies have understood short texts and classified them into a larger hierarchy using class taxonomy [30], metadata [45], and hyperbolic embedding [2], which can be applied to HS prediction. Some studies have also utilized image information for classification [19]. However, most approaches have focused on the classification itself, and lack any explanation.

2.3 Interpretable Text Classification

A large number of studies have classified text datasets using AI models. Recent studies have added interpretability to the classification model to solve the black-box nature of the deep learning model. Commonly, the classification model highlights the essential parts of a given sample to show where the model had concentrated. Some researchers provide an explanation at the sample data level and the general behavior of the model, showing frequent words and queries in learning [38]. Self-interpretable convolutional neural networks [46] have been suggested, and approaches to use max-pooling have helped interpret predictions with input tokens [4]. Other studies have utilized relations among text to extract linguistic features and to understand how a language model works [8]. Despite the ongoing efforts on interpretable text classification, an XAI model for HS code classification is unknown. Deciding HS code requires deep reasoning and it sensitively affects multiple parties; the whole process can benefit from explainable models with strong retrieval capability.

Sentence retrieval is a key module that provides explanations for the question and answering (QA) problems [32, 36]. Explanations are critical for QA tasks, where leading researchers have built numerous datasets with annotated sentences such as HotpotQA [44] and QASC [15]. Various retrieval techniques have been introduced with these datasets, such as self-attention [37], bi-attention [28], graph-based network [11], and unification-based approach [35]. However, extending these approaches is difficult when handling datasets in which ground-truth supporting facts are not given. In an unsupervised setting, alignment-based methods based on word usage and similarity have been used, such as term frequency-inverse document frequency (TF-IDF) [27] and word-mover distance [17]. The latter identifies related words when no direct matches are found between a query and a document. As the language model matures, a similarity-based pipeline can arrive at answers and supporting

sentences with high performance [12], shedding light on real-world applications with limited annotation. In the context of our research, we formulated our problem as semi-supervised sentence retrieval. We leveraged the domain expertise contained within a subset of our dataset to replicate the decision-making process of experts and enhance a model designed to assist human users. This model capitalizes on the knowledge contributed by expert annotators and builds upon the groundwork established by the previously mentioned methods.

3 DATASETS

We obtained 20 years of recent goods classification data from the Korea Customs Valuation and Classification Institute.[5] Our data span three chapters: Chapter 85 (for electrical equipments), 84 (for mechanical appliances), and 90 (for optical and photographic instruments). These chapters are known as the most challenging to classify by experts [20], because goods in these chapters have multiple functions and do not easily fit into a single HS category [25]. As a result, Chapter 85-related goods receive the most requests for expert review, accounting for nearly 17% of all classification requests in 2020 followed by Chapter 84 the second (10.1%), and Chapter 90 the fourth (6%) by Korea Customs Valuation and Classification Institute.

Furthermore, goods in these three categories share similar descriptions.[6] The task is a multi-class classification problem since these three chapters contain 163 headings and 925 subheadings in total. According to the Institute, the average classification review period to resolve classification requests in Chapter 85 is nearly 37.2 days, 36.3 days for Chapter 84, and 33.3 days for Chapter 90, far longer than the average time required for resolving goods in other categories (taking 25.9 days on average).

Table 1. The number of cases used for item classification. Three chapters—which are often confused during classification—were used for this study. The dataset includes contentious cases that were initially withheld and later were resolved by the HS council and HS committee.

	Korean case data			International
Chapter	General	Council	Committee	case data
84	5,115	231	122	55,966
85	6,434	237	192	122,221
90	4,486	166	85	31,448

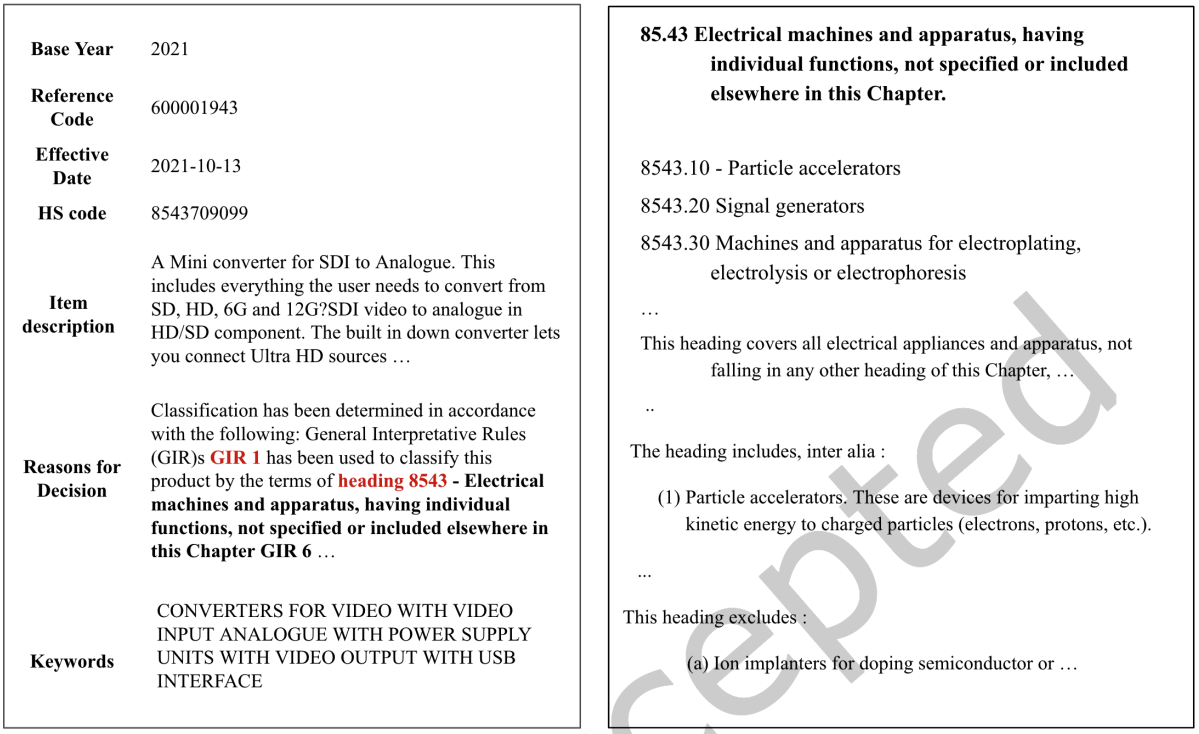
Table 1 shows the data summary we utilized in our study. The Korean data, totaling 17,068 cases, are in three levels of difficulty: 16,035 cases (or 93.9%) were resolved by the field officers at the Institute. Not all cases can be resolved by field officers, and some are moved up to be resolved at the HS council accounting for 634 (or 3.7%) of the studied cases. The most challenging cases that remain unresolved at this level are escalated to the HS committee to receive a final decision, accounting for 399 (or 2.3%) of the studied cases. We obtained the detailed supporting facts and decisions for these 1,033 contentious cases and used such information for training data.

In addition to the Korean data, International cases from 50 countries were included. The HS classification is a six-digit standard, called a subheading, for classifying globally traded products. The Institute provided international cases to use as additional training resources. We translated them into one language (i.e., Korean) and used them together to train the model. We used the first six-digits of the HS code (i.e., subheading) as it is internationally standardized. Some of the received Korean International data are made public under the international HS directory on the Customs Law Information Portal website.[7]

[5]https://www.customs.go.kr/cvnci/main.do

[6]For example, mixing units used in the sound recording are classified into heading 8543, but if it's specialized for cinematography, it belongs to heading 9010.

[7]https://unipass.customs.go.kr/clip/index.do



Base Year	2021
Reference Code	600001943
Effective Date	2021-10-13
HS code	8543709099
Item description	A Mini converter for SDI to Analogue. This includes everything the user needs to convert from SD, HD, 6G and 12G?SDI video to analogue in HD/SD component. The built in down converter lets you connect Ultra HD sources …
Reasons for Decision	Classification has been determined in accordance with the following: General Interpretative Rules (GIR)s **GIR 1** has been used to classify this product by the terms of **heading 8543 - Electrical machines and apparatus, having individual functions, not specified or included elsewhere in this Chapter GIR 6** …
Keywords	CONVERTERS FOR VIDEO WITH VIDEO INPUT ANALOGUE WITH POWER SUPPLY UNITS WITH VIDEO OUTPUT WITH USB INTERFACE

(a) Decision case sample (b) HS manual sample

Fig. 4. Sample decision data and the relevant HS manual. In (a), the given item is classified as 8543709099, 'Others in Other machines and apparatus' by H.M. Revenue and Customs. Base year, reference code, effective date, HS code, item description, reasons for the decision, and keywords are given in this example case. In (b), the manual provides the characteristics and standards of each heading (8543 here) in detail, and it includes a one-liner description of every subheading.

Figure 4 (a) shows a sample classification decision from the studied data, along with the matching HS manual. The GIR in the figure represents the General Interpretative Rules of the HS, consisting of six principles. The first principle (i.e., GIR 1) states that classification shall be determined according to the terms of the headings and any relative section or chapter notes. This example shows that the decision for HS code 854370XXXX was assigned because the description of the goods matched the guidelines for goods in heading 8543. Note, that our goal is to build an AI model that suggests the top 6 digits of the HS code (i.e., up to the subheading level).[8]

As mentioned earlier, the HS codes comprise 1,224 headings within 97 Chapters, arranged in 21 sections of the manual. Figure 4 (b) shows an example heading level from the HS manual, which starts from a heading description and is followed by a subheading. The HS manual also includes an explanation of the heading and important terminologies. In addition, it gives a list of items the heading includes and excludes.

4 INTERPRETABLE PRODUCT CLASSIFICATION MODEL

We now present an explainable product classification model. Our model takes the goods description as input and suggests the appropriate subheading (HS6 or first six digits) candidates along with some evidence. As evidence,

[8]The remaining digits describe information such as the color, shape, and material of the goods and could be determined by inspection easily.

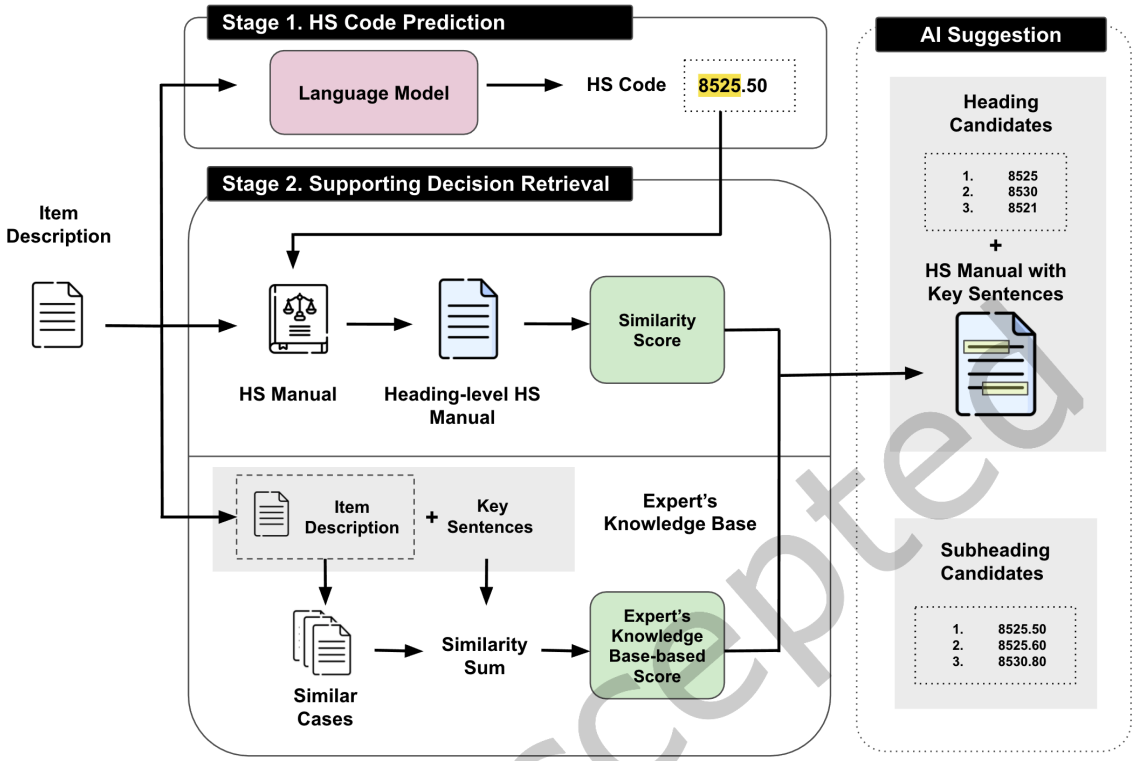


Fig. 5. The illustration of the proposed HS code classification supporting model. Stage 1 uses a language model and predicts the HS code of the goods. Stage 2 retrieves key sentences as supporting evidence using two measures: similarity (text similarity with HS manual) and expert knowledge (similar cases from the precedent). Top-3 suggestions at the subheading level (i.e., first six digits) are given in this example.

we will show relevant sentences in the HS manual that could support the decision. Figure 5 illustrates the flow of the model, which is divided into two stages: HS code prediction and supporting decision retrieval. We describe each step in detail.

4.1 Stage 1 : HS Code Prediction

Stage 1 uses a language model and predicts the HS code of the goods. Let $\mathcal{D} = \{D_1, \cdots, D_N\}$ be a collection of decision cases, where each case $D_i \in \mathcal{D}$ is a pair of the item description $\mathbf{x}_i$ and its one-hot encoded heading label $\mathbf{y}_i$. After translating all of the goods description into a common language (*i.e., Korean, for example*), we use a language model as a description encoder, e_θ, to map a sequence of words $\mathbf{x}_i$ into embedding space $\mathbb{R}^d$. Item embedding $e_\theta(\mathbf{x}_i)$ goes through the classification head, and the model is trained to minimize the loss $\mathcal{L}$ between true probability $\mathbf{y}_i$ and predicted probability $\hat{\mathbf{y}}_i = e_\theta(\mathbf{x}_i) \cdot W$, where $W \in \mathbb{R}^{d \times \dim(\mathbf{y}_i)}$ is a trainable weight matrix of the classification head. Following the rule of assigning a single HS code to each product, we regard the problem as a multiclass classification and minimize the categorical cross-entropy loss H.

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}} H(\mathbf{y}_i, \hat{\mathbf{y}}_i) \tag{1}$$

4.2 Stage 2 : Supporting Decision Retrieval

Stage 2 then retrieves key sentences as supporting evidence. To identify informative sentences, we consider two measures. One is the text similarity between the goods description and the HS manual. Second is the knowledge accumulated by experts based on previous decisions. We describe these ideas more formally below.

Let $\mathcal{M} = \{M_1, \cdots, M_K\}$ be a collection of sentences in a heading-level HS manual, find a set of relevant sentences $S_i = \{M_j\}$ with given goods description $\mathbf{x}_i$. To measure the relevancy between $\mathbf{x}_i$ and the HS manual sentence M_k, we define a relevance score $s(\mathbf{x}_i, M_k)$ as follows:

$$s(\mathbf{x}_i, M_k) = s_s(\mathbf{x}_i, M_k) + \lambda s_e(\mathbf{x}_i, M_k), \tag{2}$$

where $s_s(\mathbf{x}_i, M_k)$ is the text similarity score between two inputs, and $s_e(\mathbf{x}_i, M_k)$ is a score referring to sentences (supporting facts) written by experts to classify previous cases. λ regulates the contribution between two values. The sentences with the highest score $s(\mathbf{x}_i, M_k)$ forms S_i. The following sections explain each score in detail.

• **Text Similarity.** The first score $s_s(\mathbf{x}_i, M_k)$ is the text similarity between the sentence M_k and the given item description $\mathbf{x}_i$. Inspired by the AIR [43] model, the text similarity measures the alignment between words in M_k and $\mathbf{x}_i$. The score is high if the item description contains more words specialized in the category. $s_s(\mathbf{x}_i, M_k)$ is defined as follows:

$$s_s(\mathbf{x}_i, M_k) = \sum_{l=1}^{|\mathbf{x}_i|} idf(d_l) \cdot align(d_l, M_k), \tag{3}$$

$$align(d_l, M_k) = \max_{t=1}^{|M_k|} CosSim(d_l, m_t), \tag{4}$$

where d_l and m_t are the l^{th} and t^{th} terms of $\mathbf{x}_i$ and M_k, respectively. The cosine similarity (*CosSim*) is derived by embeddings of the two inputs from the trained language model in Stage 1, and $idf(d_l)$ is the inverse document frequency (IDF) of the word d_l.

• **Expert Knowledge.** The second score $s_e(\mathbf{x}_i, M_k)$ involves selecting key sentences based on past decisions. In certain intricate cases, pertinent information labeled as 'Reasons for Decision' was extracted from experts' input, as depicted in Figure 4. These records pertained to the resolution of contentious cases undertaken by the HS Committee and HS Council. We formed a knowledge base $\mathcal{KB} = \{(\mathbf{x}_1^{\mathcal{KB}}, E_1^{\mathcal{KB}}), \cdots, (\mathbf{x}_m^{\mathcal{KB}}, E_m^{\mathcal{KB}})\}$ by aggregating each contentious case $\mathbf{x}_j^{\mathcal{KB}}$ and its supporting facts $E_j^{\mathcal{KB}} = \{M_1, \cdots, M_a\}$. m is the number of cases in the knowledge base $\mathcal{KB}$ and a is the number of sentences for the j-th case. For instance, in the context of a classification case illustrated in Figure 4, the item description assumes the role of $\mathbf{x}_j$, and each quoted sentence, exemplified by phrases such as 'heading 8543-Electrical machines and apparatus, having individual functions, not specified or included elsewhere in this Chapter.' is represented as M_k in $E_j^{\mathcal{KB}}$. First, we pick the most relevant k_{case} cases $S^{\mathcal{KB}}(\mathbf{x}_i)$ with a given item description $\mathbf{x}_i$.

$$S^{\mathcal{KB}}(\mathbf{x}_i) = TopK_{x_j^{\mathcal{KB}} \in \mathcal{KB}}(CosSim(\mathbf{x}_i, \mathbf{x}_j^{\mathcal{KB}}), k_{case}), \tag{5}$$

where *CosSim* is the cosine similarity between two input embeddings and $TopK$ returns the most relevant k_{case} pairs in $\mathcal{KB}$ with high similarity values. Once $S^{\mathcal{KB}}(\mathbf{x}_i)$ is decided, the $\mathcal{KB}$-based similarity score $s_e(\mathbf{x}_i, M_k)$ is defined as follows:

$$s_e(\mathbf{x}_i, M_k) = \sum_{(x_j^{\mathcal{KB}}, E_j^{\mathcal{KB}}) \in S^{\mathcal{KB}}(\mathbf{x}_i)} CosSim(\mathbf{x}_i, \mathbf{x}_j^{\mathcal{KB}}) \mathbb{1}_{M_k \in E_j^{\mathcal{KB}}}. \tag{6}$$

This score represents the summation of cosine similarity values between the provided input x_i and $x_j^{\mathcal{KB}}$ within the knowledge base, contingent upon the presence of M_k in the knowledge base. For instance, if the sentence 'heading 8543-Electrical machines … in this Chapter,' is quoted by experts in the top-k most similar previous cases, it will accrue a higher score in this context.

5 QUANTITATIVE EVALUATION

We tested the feasibility of the proposed AI model using extensive quantitative and qualitative experiments. The quantitative evaluation focused on model accuracy compared to alternative methods and the quality of the suggestions.

5.1 Experimental Setting

We split the data into non-overlapping training, testing, and validation sets. In doing so, we tried to preserve the time order of the data. The first 201,435 cases were used for training the model. The most recent 5,000 cases were used for testing (including 4,324 international and 676 Korean cases). The next latest 5,000 cases were used for the validation set (4,739 international cases and 261 Korean cases) for hyperparameter tuning.

Since the AI model is intended to be used as a supporting tool, we gave the participating human inspectors multiple suggestions to choose from. Algorithmic suggestions were given for the top-k choices by accuracy for $k = 1, 3, 5$. In the retrieved sentence case study analysis, we measured recall and precision to evaluate the quality of the supporting factors.

For the baseline, we used a long short-term memory (LSTM)-based model. The LSTM-based model was a winning model in the product categorization competition in Daum shopping,[9] which has a similar setting to our problem: predict the detailed category of e-commerce products using their descriptions. The model utilizes LSTM networks to obtain embedding from tokenized input texts.

Our model was implemented over three backbone language models: KoBERT [14], KoELECTRA [7], and KLUE-RoBERTa [26] for the experiment. We used open-sourced implementations of KoBERT-base[10], KoELECTRA-base[11], and KLUE-RoBERTa-base[12]. The language models were trained for 100 epochs and evaluated when the validation accuracy was the highest. The embedding size of the language model was set to 768. We set the contribution regulating parameter λ (Equation (3)) to 0.3, and the number of similar cases k_{case} (Equation (7)) to 10 for sentence retrieval. The language model training took 40 hours and data preparation for sentence retrieval took 50 hours on an NVIDIA TITAN Xp. Inference and retrieval took less than 30 seconds.

5.2 Heading and Subheading Accuracy

Table 2 shows the top-k accuracy of the LSTM baseline and our model with different language models. Top-k accuracy was tested in both heading (HS4, first four digits) and subheading (HS6, first six digits). Our model with the KLUE-RoBERTa backbone network suggests top-3 candidates with 95.5% accuracy when classifying 163 headings, and 93.98% accuracy for 925 subheadings. These results show that our AI model performed better than the LSTM-based winning model. The table also shows that our model will improve when a more powerful backbone network is used. Note that KoBERT uses 92M parameters, taking nearly 30 minutes to train one epoch. KoELECTRA uses 113M parameters and took 24 minutes for training. KLUE-RoBERTa uses 336M parameters and takes 70 minutes to train each epoch.

[9]https://github.com/lime-robot/product-categories-classification
[10]https://github.com/SKTBrain/KoBERT
[11]https://github.com/monologg/KoELECTRA
[12]https://github.com/KLUE-benchmark/KLUE

Table 2. Classification accuracy for heading (HS4) and subheading (HS6). The result shows that the top-3 suggestions made with three language models have an accuracy of about 90% in classifying 925 subheadings.

	HS4			HS6		
Model / Top-k accuracy	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
LSTM-based	50.96	65.88	72.10	36.02	53.46	62.00
KoBERT	84.51	91.07	92.77	78.88	87.70	90.06
KoELECTRA	87.48	93.42	94.92	83.22	90.99	92.88
KLUE-RoBERTa	**89.24**	**95.50**	**96.49**	**86.06**	**93.98**	**95.25**

Table 3. Classification accuracy for contentious cases. Compared to the challenging cases escalated to the HS Committee and Council, the AI model shows outstanding performance for general HS inquiries. We note that obvious cases are not requested for classification, but only the ones that exporters and importers cannot resolve are submitted to the Institute, and are handled as a General case.

	HS4			HS6		
Dataset / Top-k accuracy	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
Committee & Council	63.04	84.70	89.13	58.70	73.17	84.78
General	89.36	95.60	96.56	86.31	94.17	95.35

In addition, we measured the performance of contentious cases in the test dataset that the HS Committee and HS Council resolved. The KLUE-RoBERTa model was used for this experiment. Table 3 shows that like human experts the AI model also had difficulty predicting the correct answer (i.e., a substantial drop in top-1 accuracy). Still, it provides helpful information with reasonably high top-3 and top-5 accuracy.

5.3 Retrieved Key Sentences

We mimicked the existing consulting documents in generating the supporting evidence for each suggestion. When field officers generate consulting documents, they quote sentences from the HS manual that give strong support for their ultimate decision. Because this document was generated manually, it had a common structure yet varied in content. Some documents also mentioned competing HS codes that were considered in the decision.

To evaluate the quality of the automatic evidence that our AI model generated, we took small samples of the consulting documents and compared them with the AI-generated ones. We compared quoted sentences from experts and retrieved sentences from our algorithm at the sentence level.

Table 4 shows an example of the evidence generated by the field officer (on top) and by the algorithm (bottom). This particular case of heading '8472' had 75 original sentences in the HS manual. We compared which sentences were highlighted by the algorithm and which ones were chosen as evidence by human experts. This test had a recall and precision of 0.75 each. The first sentence retrieved from our model matches the second sentence from the expert. Similarly, the second sentence from our model matches the third one from the expert, and the third sentence matches the fourth one from the expert.

Table 4. Comparison of sentences written by experts and sentences retrieved by our model in a sample case. Three out of four sentences retrieved by our model were equivalent to the expert's evidence sentences.

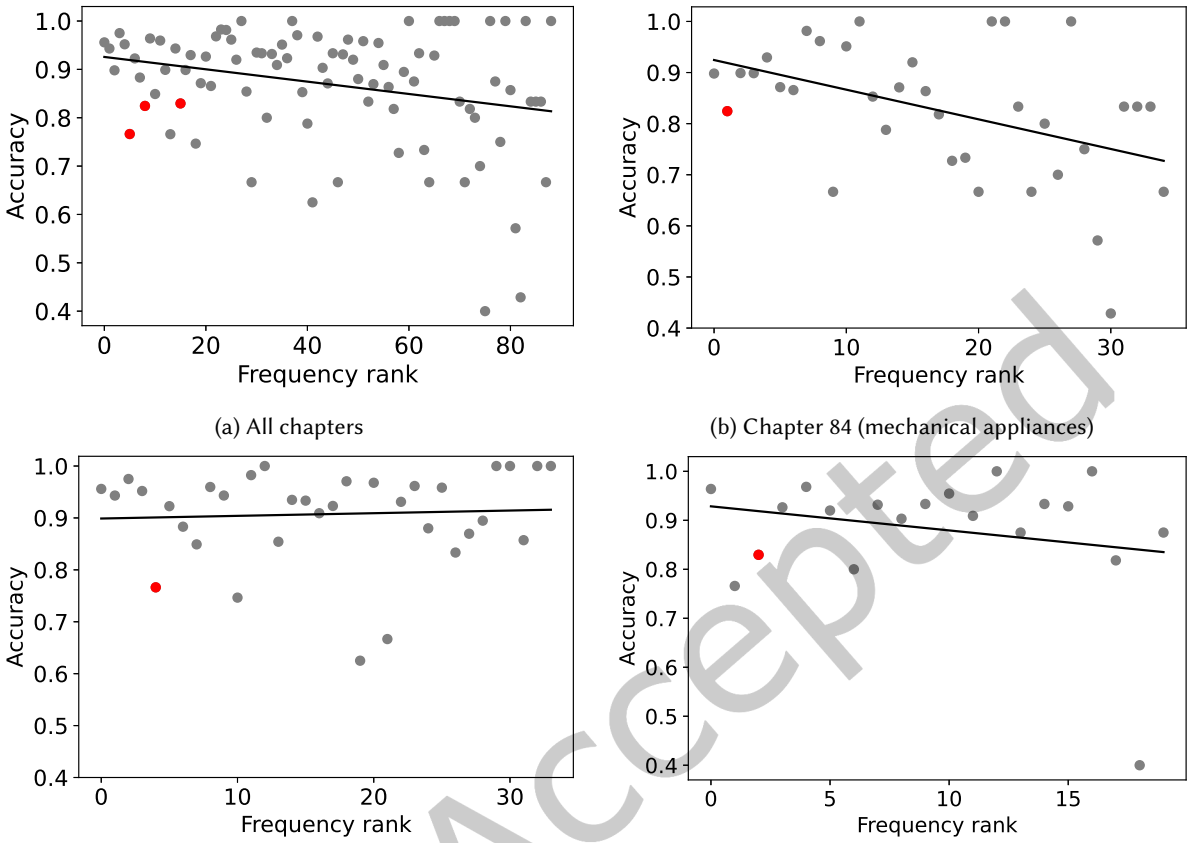
Reasons for decision by experts
1. 84.72 Other office machines (for example, hectograph or stencil duplicating machines, addressing machines, automatic banknote dispensers, coin-sorting machines, coin-counting or wrapping machines, pencil-sharpening machines, perforating or stapling machines).
2. This heading covers all office machines not covered by the preceding two headings or more specifically by any other heading of the Nomenclature.
3. The term "office machines" is to be taken in a wide general sense to include all machines used in offices, shops, factories, workshops, schools, railway stations, hotels, etc., for doing "office work" (i.e., work concerning the writing, recording, sorting, filing, etc., of correspondence, documents, forms, records, accounts, etc.).
4. Office machines are classified here only if they have a base for fixing or for placing on a table, desk, etc. The heading does not cover the hand tools, not having such a base, of Chapter 82.
Supporting facts found by our model
1. This heading covers all office machines not covered by the preceding two headings or more specifically by any other heading of the Nomenclature. → **Eqv. to (2)**
2. The term "office machines" is to be taken in a wide general sense to include all machines used in offices, shops, factories, workshops, schools, railway stations, hotels, etc., for doing "office work" (i.e., work concerning the writing, recording, sorting, filing, etc., of correspondence, documents, forms, records, accounts, etc.) → **Eqv. to (3)**
3. Office machines are classified here only if they have a base for fixing or for placing on a table, desk, etc. The heading does not cover the hand tools, not having such a base, of Chapter 82. → **Eqv. to (4)**
4. Automatic banknote dispensers, operating in conjunction with an automatic data processing machine, whether on line or off line.

To test the quality of the evidential sentences, we also obtained 15 new contentious cases listed on the Customs Law Information Portal website. None of these cases had been used for training our model. Our AI model generated suggestions for these complex cases and retrieved seven sentences for these cases, based on feedback from the field officers, many of whom indicated seven as the most preferred number of sentences to view.
We report the recall scores between the automatically highlighted sentences and those indicated by a human expert. This evaluation focuses on the recall value because the AI model should not miss any key evidential sentences that human experts consider important. The precision metric is excluded since it can differ depending on the number of sentences to view. The average recall was 0.69 for the 15 cases, indicating that the AI model chose the same evidential sentence as humans with a probability of 69%.

5.4 Relationship between Frequency and Accuracy

Figure 6 shows the accuracy of each heading from the ELECTRA model. This figure depicts the trend in prediction accuracy as a function of heading frequency in the dataset. The negative slope of the fitted lines in Chapters 84 and 90 indicates that the AI model provided higher prediction accuracy for more frequently appearing goods. This is a desirable trait, as top-ranked items take up a disproportionately larger part of the entire data. However, Chapter 85 does not show the same characteristic, likely because it contains challenging cases.

Each Chapter contains a heading category named *Miscellaneous* which is used to hold a variety of cases that do not perfectly fit any given heading within the Chapter. This Miscellaneous category was one of the frequently



(a) All chapters

(b) Chapter 84 (mechanical appliances)

(c) Chapter 85 (electronic equipment)

(d) Chapter 90 (optical and photographic)

Fig. 6. The correlation between heading-level frequency and the prediction accuracy of the ELECTRA-based AI model. The AI model provided higher prediction accuracy for more frequently appearing goods in (b) Chapter 84 and (d) Chapter 90. (c) Chapter 85 goods include challenging cases and do not show the same desirable pattern. The red dots for each Chapter indicate the *Miscellaneous* category, which shows below-average performance.

requested categories for classification, yet as one may expect, the AI model also gave below-average prediction accuracy.

6 QUALITATIVE EVALUATION

We next tested the usability of the AI model via a survey. To determine how field officers perceived an AI assistant tool, we built a prototype service that could be accessed via the Web. The prototype system did not require any understanding of the model's inner workings. Our collaborating partners at the Korea Customs Service (KCS) helped recruit field officers at the Korea Customs Valuation and Classification Institute (Number of participants, N=32), who tested the efficacy of the prototype AI model. Customs field officers with varying career experiences participated in our usability survey. Below, we describe key findings and feedback from this survey study.

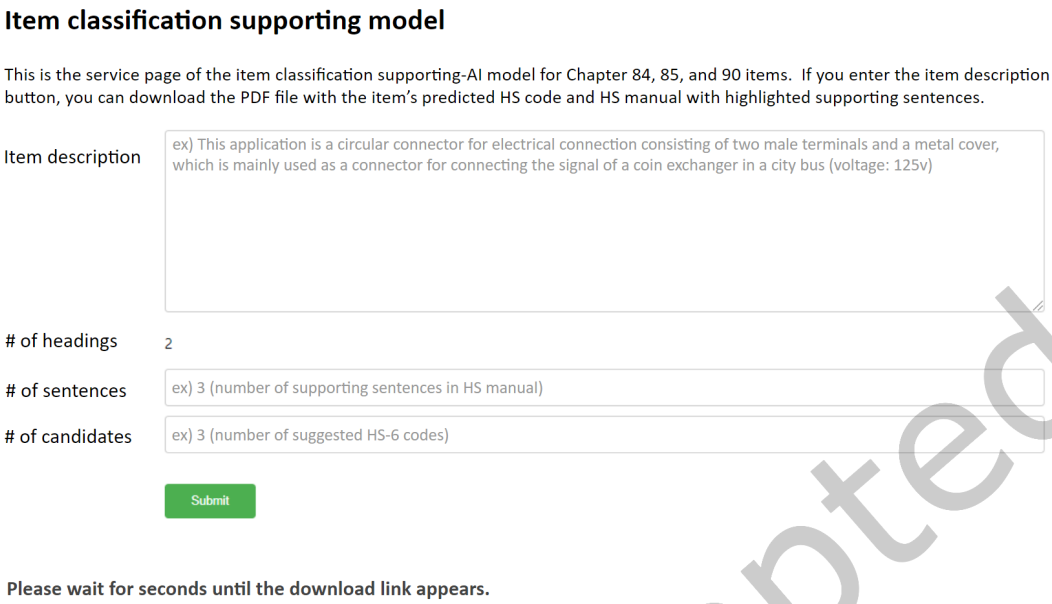


Fig. 7. Web interface of the prototype AI model tested by field officers

6.1 Prototype Model

Figure 7 is a screen snapshot of the prototype Web service, which was designed to assist field officers in their daily classification tasks. The page allowed easy adjustment of the number of candidates and the length of the evidence sentences to be shown on the screen. Officers could type in or copy text descriptions of the goods they needed to classify, to start a new query. The top suggestions and the corresponding evidential sentences were given in a PDF file.

Figure 8 is an example output. The output document consists of three parts: 1) entered item description, 2) candidate 4-digit heading with supporting sentences from the HS manual, 3) candidate 6-digit subheadings. The document includes the entire text of the candidate headings' HS manual, highlighting important sentences. We designed the output document with two goals in mind. The first was to mimic the HS Council and HS Committee's consulting documents which are produced for disputed cases. Like those documents, we quoted sentences from the HS manual that were the most characteristic of the suggestion. The second goal was to improve convenience for the field officers. Typically, the officers would refer to the HS manual if suggestions were relevant. Hence, we provided a complete version of the corresponding HS manual and highlighted the evidential sentences in red text. We also provided a calibrated prediction score for each candidate to indicate model confidence. Temperature scaling [13] was applied on $\hat{\mathbf{y}}_i$ to adjust the range of values.

6.2 Qualification Analysis

Our partners at the Korea Customs Service (KCS) helped recruit field officers of varying work experience to test the prototype. Participants were given two weeks to experience the prototype web page and received suggestions using the AI model in real-world situations. The usability survey was designed to evaluate the quality of the model's classification support under real-world situations, and 32 officers who tested the service responded to the

1 Item description

*Overview of the requested item. - This item is installed in construction equipment (excavator, wheel loader, etc.) and is used to transmit and receive location information through GPS antennas and driving information (operation time, etc.) to and from communication servers through 3G antennas. *Components and functions of the requested item. - 3G antenna Rod: Data transmission/reception manager. - GPS antenna: Receipt of location information, etc. - GPS connection cable: connected to the GPS input terminal of the RMCU. - 3G connection cable: connected to the data transmission/reception input terminal of the RMCU.

2 Candidate heading : 8517 (score : 0.612)

85.17- Telephone sets, including telephones for cellular networks or for other wireless networks; other apparatus for the transmission or reception of voice, images or other data, including apparatus for communication in a wired or wireless network (such as a local or wide area network), other than transmission or reception apparatus of heading 84.43, 85.25, 85.27 or 85.28 (+).

- Telephone sets, including telephones for cellular networks or for other wireless networks :

8517.11 - Line telephone sets with cordless handsets

8517.12 - Telephones for cellular networks or for other wireless networks

8517.18 - Other

- Other apparatus for transmission or reception of voice, images or other data, including apparatus for communication in a wired or wireless network (such as a local or wide area network) :

8517.61 - Base stations

⋮

(ij) Insulated electric wire, cable, etc., as well as optical fibre cables, made up of individually sheathed fibres, whether or not fitted with connectors, including cords with plugs for switchboards (heading 85.44).

(k) Telecommunication satellites (heading 88.02)

(l) Telephone call registers and counters (Chapter 90).

(m) Carrier-current and other transmitters and receivers which form a single unit with analogue or digital telemetering instruments or apparatus, or which, together with the latter, constitute a functional unit within the meaning of Note 3 to Chapter 90 (Chapter 90).

(n) Calculographs (time recorders) (heading 91.06).

(o) Monopods, bipods, tripods and similar articles (heading 96.20).

...

3 Top-3 Candidate Subheadings

3.1 851770 (score : 0.942)

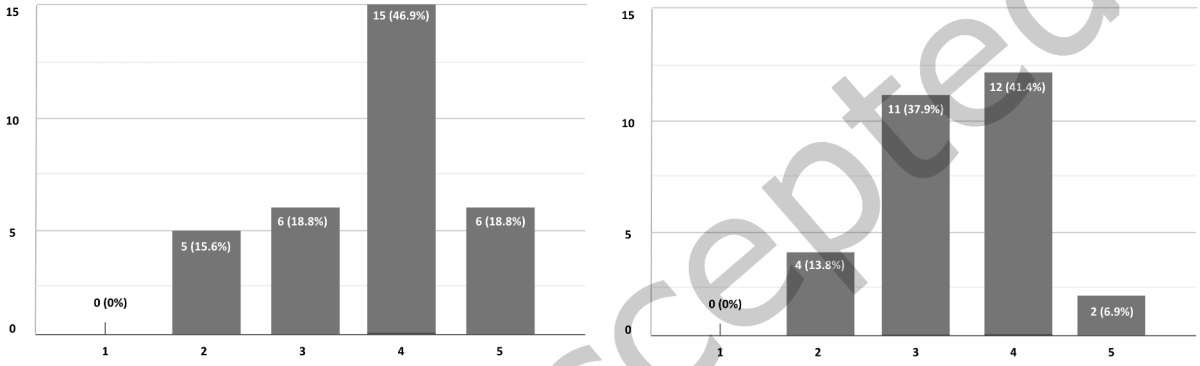
3.2 851762 (score : 0.051)

3.3 851711 (score : 0.003)

Fig. 8. Prototype model results provided to field officers. The entered item description, predicted heading, candidate heading's HS manual with supporting sentences, and candidate codes (subheading) are given.

survey. The survey included the following questions, including both Likert-scale and open-ended ones. Survey responses were anonymous:

(1) How helpful was the AI suggestion? (Likert-scale was 1: not very helpful, 2: not helpful, 3: neutral, 4: helpful, 5: very helpful.)
(2) Please describe if the assistant tool was helpful for your task, and if so, how?
(3) How accurate was the AI suggestion between 1 and 5? (Likert-scale was 1: not very accurate, 2: not accurate, 3: neutral, 4: accurate, 5: very accurate.)
(4) Please describe your thoughts on the accuracy of the AI suggestions.
(5) How many candidate suggestions and evidence sentences would you like to see?
(6) How long have you worked in the customs field and on the HS code classification task?



(a) Helpfulness (b) Accuracy

Fig. 9. Results of the usability survey. (a) Distribution of helpfulness between 1 (not very helpful) and 5 (very helpful). (b) Distribution of accuracy between 1 (not very accurate) and 5 (very accurate).

• **Helpfulness and Accuracy.** Survey participants queried the system, with distinct goods and numerous variations in the number of candidate suggestions and evidential sentences as depicted in Figure 7. Figure 9 shows the distribution of the helpfulness and accuracy responses.

65.7% of the participants answered that the AI suggestions were helpful, with a score of four or five. Accuracy also showed a positive response overall. We found that more than 85% gave a score of three or more for accuracy. In addition, there was a tendency that the participant who responded as 'helpful' also rated higher accuracy. The Pearson correlation between the helpfulness score and accuracy was 0.82, and the Spearman correlation was 0.79.

• **Analysis by Career Experience.** Based on the final survey question about the career experience at customs and the HS classification task, we revisited the responses on helpfulness and accuracy. We identified five respondents each who had the longest work experience at customs as Group A, the shortest work experience at customs as Group B, the longest work experience at HS classification task as Group C, and the shortest work experience at HS classification task as Group D. The respondents in Group A on average had spent over 22 years in customs service, and Group B spent fewer than six years. Group C had worked on average longer than seven years on the classification task, and Group D had spent less than a year on the task.

Figure 10 shows the score of helpfulness answered by four groups. The top plots comparing Group A and Group B indicate that less experienced officers with a shorter working period in customs service found the AI assistant tool substantially more helpful. Four out of five respondents gave a score of 4, and one gave a score of 5.

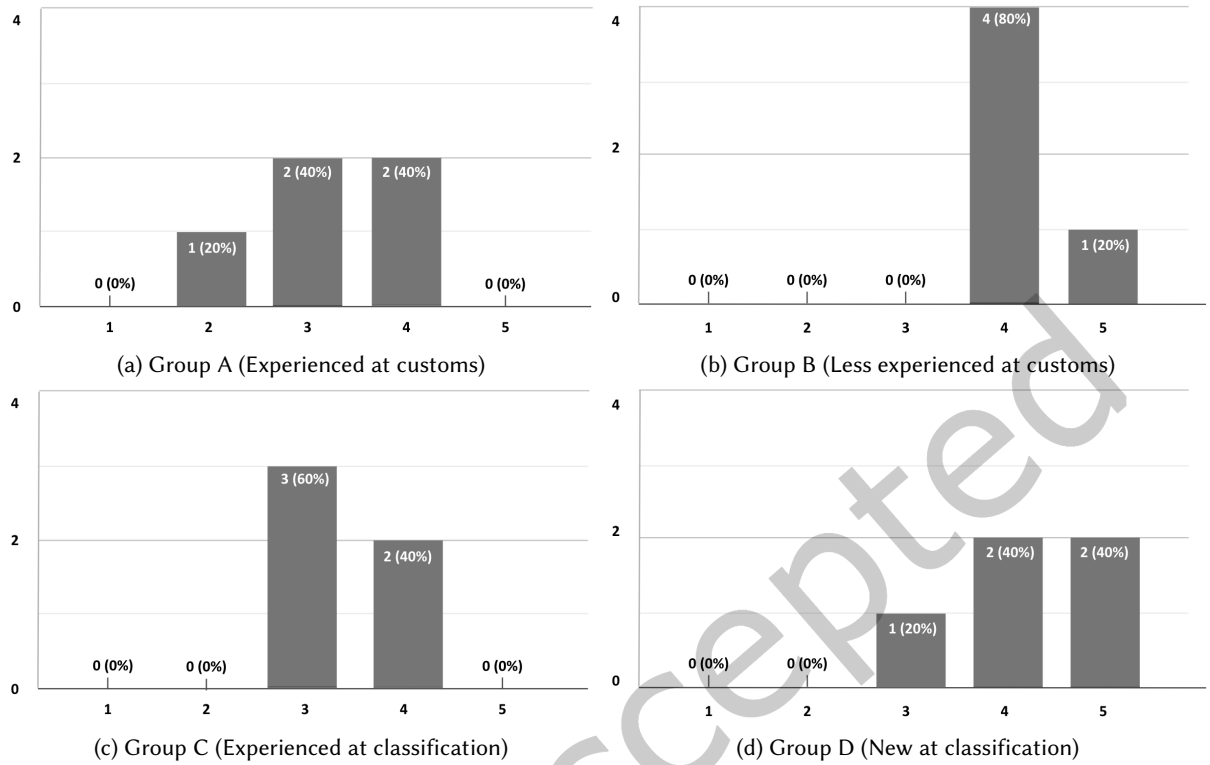


Fig. 10. Distribution of helpfulness scores by work experience. (a) Group A has the longest work experience at customs service, (b) Group B has the shortest work experience at customs service, (c) Group C has the longest work experience at classification task, and (d) Group D has the shortest work experience at classification task.

Even among the experienced officers with an average of 22 years in the career, two out of five respondents gave a score of 4 in helpfulness.

The bottom two plots comparing Group C and Group D reinforced our finding that less experienced officers tended to find the AI tool more helpful. Among the least experienced officers who had been on the classification task for less than a year, two out of five gave a helpfulness score of 5.

These findings suggest that AI models can effectively support novice customs officers in their tasks. The feedback received from senior officers indicates a desire for more precise and comprehensive information beyond what they already know from the system, resulting in relatively lower scores. This can be interpreted as an indication that the system possesses similar knowledge and classification abilities as the experienced officers themselves.

• **Open-Ended Feedback.** Next, we examined the open-ended feedback to understand what aspect of the AI tool was considered the most (or the least) helpful (or accurate). The overall feedback was positive, and some respondents were surprised with the level of accuracy the model could achieve on the challenging task. Here are some quotes from the feedback.

One respondent shared a perspective that the AI tool acted as a 'second eye' for decision making:

"*AI suggestions were helpful because I could compare my decision with them.*" (P1)

Other participants also resonated with this perspective, since validation is critical in HS classification, knowing the decision directly affects the tariff rates and has a huge financial impact on importers and exporters. Field officers valued that AI could act as a validation tool for their own decisions. Similar to increasing the accuracy of classification, there was also a mention that the tool helped reduce potential errors.

"*The model gave suggestions that I could have missed. I found this very helpful.*" (P13)

Another common perspective shared was the tool's ability to reduce the time needed for initial investigation. Many field officers started by listing a wide pool of candidate HS codes, and then filtering down to a smaller set of candidate decisions. The AI tool was able to assist this initial investigation by introducing a large number of candidate suggestions. Here are some relevant quotes:

"*I think this tool can help shorten my screening time.*" (P14)

"*The model helped me make quick decisions.*" (P29)

When it came to helpfulness, several participants responded that candidate codes and the evidential sentences highlighted from the HS manual were useful in making a decision:

"*The algorithmic suggestions and the supporting sentences helped me reduce the candidate pool to review.*" (P7)

"*I found the snippet of HS manual given with candidate suggestions helpful. I could concentrate on the model's reasoning sentences and references.*" (P3)

Some officers mentioned the potential for the tool to be used as an educational tool, as it could give an overview of the classification task. The feedback here could also be appreciated together with the high accuracy and helpfulness scores shown for the less experienced officers. Here are some relevant quotes:

"*The supporting model gave a rough idea of final decision I had to make.*" (P20)

"*Since the model shows the candidates, it can be helpful to educate new workers who have short working experience and expertise in the classification task.*" (P12)

Last, related to how many candidates and supporting evidence individuals wanted to see, participants were most comfortable with the visual setting when the AI model provided three candidate subheadings and seven evidential sentences. The exact average was 3.17 subheadings and 6.74 sentences each. Notably, when examining the preferences of junior officers (the 10 least experienced officers) and senior officers (the 10 most experienced officers), it was observed that junior officers favored 3.0 candidates and 5.8 sentences, while senior officers leaned towards 3.5 candidates and 5.4 sentences. However, despite these variations, the differences in responses between the two groups were not substantial. It is noteworthy that both junior and senior officers exhibited a similar viewpoint regarding the desired number of candidates and sentences.

7 DISCUSSION

7.1 Toward Interpretable Results

Interpretability is critical in many high-risk scenarios like HS classification. We provided a confidence score for each HS code candidate, which provides additional information to help the user judge whether or not a candidate was valid. Although the score is tuned by temperature scaling, the range of the top-k confidence score is quite different for each input item. Careful calibration is required to encourage customs officers to use this value as a reference in decision-making.

Another way to increase interpretability is to visualize the part of the item description related to each subheading candidate; then, customs officers can concentrate on the selected part and decide whether to consider second and third candidates to review. In addition, key sentences should relate to the subheading characteristics so that the final form of model output resembles the reports written by experts. Creating an organized document that

explains the relations among prediction, description, and HS manual can reduce the effort required for HS code classification.

These records pertained to the resolution of contentious cases undertaken by the HS Committee and HS Council

As this work demonstrated, there is a substantial resemblance in how customs experts decide on the HS codes with how judges decide on legal cases [41]. Moreover, deep learning models solve classification problems by finding common patterns from previous cases. As a result, past examples are the primary determinants of the AI model's decision, unlike human experts, who make decisions based on rules and manuals. Since the HS code and its manual undergo revisions every five years, previous cases cannot always be good references for recent ones. Revision can be viewed as an update of the knowledge base. This makes it essential to employ GRIs and the HS manual to prepare a credible model, which utilizes contextual information in model training based on deep linguistic understanding [39].

7.2 Challenges

• **Data Distribution Shift.** As the AI model is based on the learned knowledge from existing classification, a shift in data distribution can affect the quality of the suggestion. However, such shifts are a common part of many live systems, and this poses a challenge for the longitudinal adoption of an AI assistant model. Trade data is no exception. Figure 11 shows example headings that have been increasing in their request at the HS classification task, at the same time other requests have been decreasing over a decade. This natural change reflects the adoption of new technology in developing products. One may re-learn these changes by updating the AI model's knowledge from time to time or by attempting to continue learning new patterns.

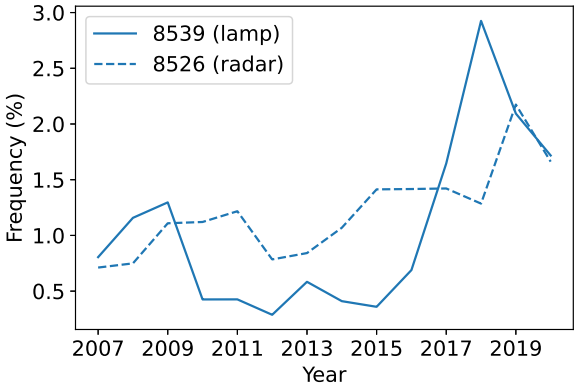
Another aspect is an update in the HS manual itself, which goes through extensive revision and updates every five years to represent technological advances better. For example, when the smartwatch was launched, there was only an HS code for smart mobile devices and a traditional watch, but not a combination. Choosing either selection will lead to different tariff rates. The same goes for many other technological advances. The World Customs Organization (WCO) considers these changes and introduces new subheadings or removes subheadings that are no longer used, etc. The most recent release of the HS manual is the seventh edition which had 351 sets of amendments, effective from January 1, 2022[13]. In this new edition, subheading *8517.13* representing '*Smartphones*' newly appears. Previously, smartphones were usually classified into *8517.12* 'Telephones for cellular networks or other wireless networks.'

As mentioned earlier, the AI model needs to take these new changes into account and re-learn the manual. Without a specific update, the model suggestions will no longer be relevant over time [16]. Once a model has been deployed, continual monitoring of the parameter stability and model performance is required [5]. The maintainer should retrain the model whenever the system detects a distribution shift. *Continual learning* is a method that allows such model training on new data [23] and this idea can be applied to our AI model as an extension.

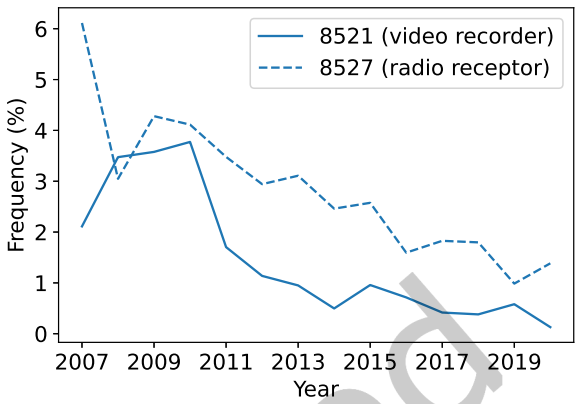
• **Data Imbalance and Vagueness.** Another challenge in building the AI model is data imbalance and vagueness. As we have shown in Figure 3, the headings and subheadings follow a skewed distribution. Some appear disproportionately more in the data, whereas others appear substantially less frequently. Together with the temporal nature of data, data imbalance could pose a challenge in predicting non-popular items.

As the final challenge, we point back to the prevalence of the *Miscellaneous* category. We have shown that this category is more difficult to predict than others, as it contains a variety of items that do not fit perfectly with other subheadings or headings. Figure 12 shows this challenge once more, indicating the popularity of the *Miscellaneous* category (marked in red color) along with their proportions in the data. This category continues to become

[13]http://www.wcoomd.org/en/topics/nomenclature/instrument-and-tools/hs-nomenclature-2022-edition.aspx



(a) Headings whose share increased over time. Heading 8539 represents sealed beam lamp units and arc lamps, and 8526 is for radar apparatus or radio remote control apparatus including GPS tracker.



(b) Headings whose share decreased over time. Heading 8521 includes video recording or reproducing apparatus, and 8527 is for reception apparatus for radio-broadcasting.

Fig. 11. Data distribution shift: Four representative headings show changes in trade patterns over time.

larger with new technological advances and will likely remain a challenging case to predict for both AI and humans.

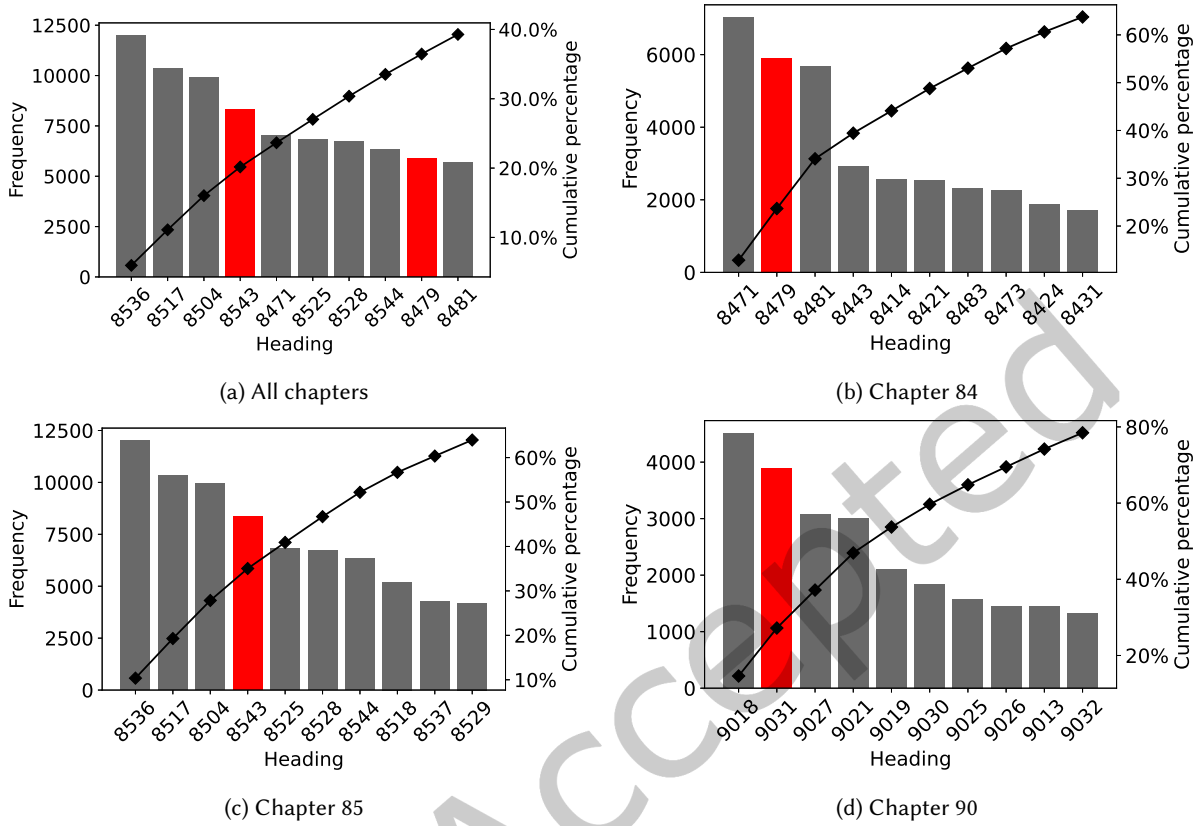
7.3 Possibility of Utilizing Large Language Models (LLMs)

In order to explore the capabilities of the latest large language models, specifically ChatGPT, we conducted a series of toy experiments to determine whether LLM could classify customs products without the need for additional training steps. Our approach involved providing ChatGPT with the General Rules for the Interpretation of the Harmonized System (GRI) and a single heading-level HS manual. Specifically, we focused on the '8533: Electrical resistors' heading, which comprises seven subheadings as sub-classes. We presented ChatGPT with five item descriptions belonging to the '8533' chapter and tasked it with predicting the appropriate subheading for each item.

Remarkably, ChatGPT successfully predicted the correct subheading for four out of the five items. However, it encountered a misclassification issue, erroneously assigning the sub-heading '8533.31' to '8533.39' due to a misunderstanding of the numerical details within the item characteristics. Despite this classification test took place with the small number of class candidates, we observed that LLM can comprehend the standards associated with each HS code in a zero-shot manner. Based on these encouraging results from our toy experiment, we believe there is potential for applying the latest LLMs to address complex classification problems in this domain.

Additionally, an intriguing aspect of ChatGPT's performance was its ability to provide explanations for its classification decisions, along with the predicted class, when the request was solely focused on classification. This interpretability feature holds promise for facilitating the handling of rapidly changing customs item classification tasks and generating understandable outcomes.

Nevertheless, it is important to acknowledge the existing limitations. To further assess the capabilities of ChatGPT, we introduced two more HS manuals for the '8534' and '8535' classes, in total 14 subheadings. Subsequently, we provided ChatGPT with five item descriptions and requested classification tasks in conditions where there



(a) All chapters (b) Chapter 84

(c) Chapter 85 (d) Chapter 90

Fig. 12. Top-10 frequent headings in each chapter. Red bars indicate *Miscellaneous* headings, which occupy a fairly large proportion.

were more candidates to choose from. In this scenario, ChatGPT achieved three correct predictions out of five, and it also exhibited a common challenge: providing plausible yet incorrect answers. Notably, it misclassified the last product as '8536.69 - Electrical apparatus for connecting electrical circuits, for a voltage exceeding 1,000 V,' which is a non-existing subheading. Given the sensitivity of customs item classification, the issues of ambiguity and data insecurity pose significant challenges when leveraging trained LLMs. Addressing these concerns represents a crucial area for future research to harness the potential of these valuable tools effectively.

8 CONCLUSIONS

This research presents an AI model for assisting with the process of HS code classification at customs offices. Using the product description and the HS manual, the model predicted the first 4-digit headings and 6-digit subheadings along with supporting facts related to the suggestion to human experts. Rather than replacing human judgment entirely, the model gave top suggestions as a guiding tool. We also had a unique opportunity to collaborate with the Korea Customs Service and test the feasibility of the AI model as a prototype service with field officers (N=32).

We expect that our work will contribute substantially in various respects. The use of this AI assistant tool by declarants can improve the initial declaration quality, thereby reducing workload at customs offices, particularly when competing HS codes are problematic for declarants and customs officials. Internally, the tool could assist customs officials in the various ways identified in the survey, for example, as an educational tool for new officials, as a validation tool for experienced officials, and as a guiding tool that helps reduce the time and effort needed to screen for candidate codes by all officials.

Our model presents the competing HS codes of the target product with its rationale, so it has great significance as an auxiliary means for product classification. Platforms require a systematic classification system to effectively expose and recommend products to users, but each product-providing company often has different standards. The platforms build and utilize hierarchical classification algorithms to maintain the consistent categorization of hundreds of millions of products. Our work can be used to advance these algorithms and classifications and facilitate their management.

Looking to the future, large language models (LLMs) offer promising solutions for addressing the complexities of customs classification tasks. LLMs possess the ability to understand the nuances of customs item descriptions through their extensive pre-training. However, it's essential to cautiously integrate LLMs into our proposed algorithm while considering the risks associated with prediction errors, particularly hallucinations. In closing, our research has introduced a robust and transparent classification model for customs goods. Future work should prioritize the effective utilization of LLMs to enhance model adoption and performance within this context.

ACKNOWLEDGMENTS

This research was supported by the Institute for Basic Science (IBS-R029-C2), NRF grant (RS-2023-00240062), and the IITP grant (RS-2023-00216011, 2019-0-01842) by the Ministry of Science and ICT in Korea. We thank Minsoo Song, Junok Kang, Sungdae Ji, Yeonsoo Choi from Korea Customs Service for their insightful discussions.

REFERENCES

[1] Fatma Altaheri and Khaled Shaalan. 2020. Exploring Machine Learning Models to Predict Harmonized System Code. In *Proc. of the European, Mediterranean, and Middle Eastern Conference on Information Systems*. 291–303.

[2] Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. 2020. Hyperbolic Interaction Model For Hierarchical Multi-Label Classification. In *proc. of the AAAI Conference on Artificial Intelligence*. 7496–7503.

[3] Xi Chen, Stefano Bromuri, and Marko van Eekelen. 2021. Neural Machine Translation for Harmonized System Codes Prediction. In *proc. of the International Conference on Machine Learning Technologies (ICMLT)*. 158–163.

[4] Hao Cheng, Xiaoqing Yang, Zang Li, Yanghua Xiao, and Yucheng Lin. 2019. Interpretable Text Classification Using CNN and Max-Pooling. *arXiv preprint arXiv:1910.11236* (2019).

[5] Huyen Chip. 2022. Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications.

[6] Ciel HS. 2022. Harmonized commodity description and coding system explanatory notes. https://www.clhs.co.kr/uploads/lawfile/404n2563.pdf. Accessed: 2022-02-27.

[7] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators. In *proc. of the International Conference on Learning Representations (ICLR)*.

[8] Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *proc. of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 306–316.

[9] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2021. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* (2021), 101666.

[10] Liya Ding, ZhenZhen Fan, and DongLiang Chen. 2015. Auto-Categorization of HS Code Using Background Net Approach. *Procedia Computer Science* 60 (2015), 1462–1471. https://doi.org/10.1016/j.procs.2015.08.224

[11] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8823–8838.

[12] Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A Simple Yet Strong Pipeline for HotpotQA. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8839–8845.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *proc. of the International Conference on Machine Learning (ICML)*. 1321–1330.
[14] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.
[15] Tushar Khot, Peter Clark, Michal Guerquin, Peter A. Jansen, and Ashish Sabharwal. 2020. QASC: A Dataset for Question Answering via Sentence Composition. In *proc. of the AAAI Conference on Artificial Intelligence*. 8082–8090.
[16] Sundong Kim, Tung duong Mai, Sungwon Han, Sungwon Park, Thi Nguyen, Jaechan So, Karandeep Singh, and Meeyoung Cha. 2022. Active Learning for Human-in-the-loop Customs Inspection. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[17] Sun Kim, Nicolas Fiorini, W John Wilbur, and Zhiyong Lu. 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of biomedical informatics* 75 (2017), 122–127.
[18] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.
[19] Dongju Lee, Gunwoo Kim, and Keunho Choi. 2020. CNN-based Recommendation Model for Classifying HS Code. *Management & Information Systems Review* 39, 3 (2020), 1–16.
[20] Eunji Lee, Sundong Kim, Sihyun Kim, Sungwon Park, Meeyoung Cha, Soyeon Jung, Suyoung Yang, Yeonsoo Choi, Sungdae Ji, Minsoo Song, and Heeja Kim. 2021. Classification of goods using text descriptions with sentences retrieval. In *proc. of the Korea Artificial Intelligence Conference (KAIA)*.
[21] Guo Li and Na Li. 2019. Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electronic Commerce Research* 19, 4 (2019), 779–800.
[22] Jeffrey Luppes, Arjen P de Vries, and Faegheh Hasibi. 2019. Classifying short text for the harmonized system with convolutional neural networks. *Radboud University* (2019).
[23] Tung-Duong Mai, Kien Hoang, Aitolkyn Baigutanova, Gaukhartas Alina, and Sundong Kim. 2021. Customs fraud detection in the presence of concept drift. In *ICDM IncrLearn Workshop*.
[24] Carolyn McKay. 2020. Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice* 32, 1 (2020), 22–39.
[25] Minkyu Park. 2019. A study on the customs classification fallacy of certain ITA goods. *Korea Trade Review* 44, 2 (2019), 189–202.
[26] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. In *proc. of Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*. 1–25.
[27] Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *proc. of the International Conference on Machine Learning (ICML)*, Vol. 242. 29–48.
[28] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *proc. of the International Conference on Learning Representations (ICLR)*.
[29] Korea Customs Service. 2018. E-commerce goods import trend. https://url.kr/4bp9ew. Accessed: 2023-07-16.
[30] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4239–4249.
[31] Tania Sourdin. 2018. Judge v Robot?: Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal* 41, 4 (2018), 1114–1133.
[32] Mokanarangan Thayaparan, Marco Valentino, Viktor Schlegel, and André Freitas. 2019. Identifying Supporting Facts for Multi-hop Question Answering with Document Graph Networks. In *proc. of the Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs)*. 42–51.
[33] The Korea Times. 2015. Smartwatch is a communication device. https://tinyurl.com/4vrfx7ef. Accessed: 2022-02-27.
[34] Santosh T.Y.S.S, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding Legal Judgment Prediction for European Court of Human Rights Cases Towards Better Alignment with Experts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1120–1138.
[35] Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. Unification-Based Reconstruction of Multi-Hop Explanations for Science Questions. In *proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
[36] Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *proc. of the Conference on Computational Natural Language Learning (CoNLL)*. 696–707.

[37] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 189–198.
[38] Zhengyang Wang, Xia Hu, and Shuiwang Ji. 2020. iCapsNets: Towards Interpretable Capsule Networks for Text Classification. *https://arxiv.org/abs/2006.00075* (2020).
[39] Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable NLP. In *proc. of Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*.
[40] Wikipedia. 2022. General rules for the interpretation of the harmonized system. https://en.wikipedia.org/wiki/General_Rules_for_the_Interpretation_of_the_Harmonized_System. Accessed: 2022-02-27.
[41] Christoph Winter. forthcoming. The Challenges of Artificial Judicial Decision Making for Liberal Democracy. In *Judicial Decision-making: Integrating Empirical and Theoretical Perspectives*.
[42] World Customs Organization. 2018. HS compendium – The harmonized system, a universal language for international trade. https://tinyurl.com/ycr259ty. Accessed: 2022-02-27.
[43] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering. In *proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 4514–4525.
[44] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2369–2380.
[45] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In *proc. of the Web Conference (WWW)*. 3246–3257.
[46] Wei Zhao, Rahul Singh, Tarun Joshi, Agus Sudjianto, and Vijayan N Nair. 2021. Self-interpretable Convolutional Neural Networks for Text Classification. *arXiv preprint arXiv:2105.08589* (2021).