The use of machine learning to identify the correctness of HS Code for the customs import declarations

Chen, Hao; Van Rijnsoever, Ben; Molenhuis, Marcel; van Dijk, Dennis; Tan, Y.; Rukanova, B.D.

DOI
10.1109/DSAA53316.2021.9564203

Publication date
2021

Document Version
Final published version

Published in
IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)

The use of machine learning to identify the correctness of HS Code for the customs import declarations

Hao Chen
Innovation Exchange
IBM Ireland
Dublin, Ireland
marc.chenhao@gmail.com

Ben van Rijnsoever
GBS Global
IBM Netherlands
Amsterdam, Netherlands
b.rijnsoever@nl.ibm.com

Marcel Molenhuis
Secretaris CoördinatieGroep Innovatie
Dutch Customs Administration
Rotterdam, Netherlands
jm.molenhuis@douane.nl

Dennis van Dijk
Innovatie en Business Analytics
Dutch Customs Administration
Rotterdam, Netherlands
d_van_dijk@belastingdienst.nl

Yao-hua Tan
Department of Technology, Policy and Management
Delft University of Technology
Delft, Netherlands
y.tan@tudelft.nl

Boriana Rukanova
Department of Technology, Policy and Management
Delft University of Technology
Delft, Netherlands
b.d.rukanova@tudelft.nl

2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA) | 978-1-6654-2099-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/DSAA53316.2021.9564203

***Abstract*—As an increasing volume of international trade activities around the world, the amount of cross-boarder import declarations grows rapidly, resulting in an unprecedented scale of potentially fraudulent transactions, in particular false commodity code (e.g., HS Code). The incorrect HS Code will cause duty risk and adversely impact the revenue collection. Physical investigation by the customs administrations is impractical due to the substantial quantity of declarations. This paper provides an automatic approach by harnessing the power of machine learning techniques to relief the burden of customs targeting officers. We introduced a novel model based on the off-the-shelf embedding encoder to identify the correctness of HS Code without any human effort. Determining whether the HS Code is correctly matched with commodity description is a classification task, so the labelled data is typically required. However, the lack of gold standard labelled data sets in customs domain limits the development of supervised-based approach. Our model is developed by the unsupervised mechanism and trained on the unlabelled historical declaration records, which is robust and able to be smoothly adapted by the different customs administrations. Rather than typically classifying whether the HS Code is correct or not, our model predicts the score to indicate the degree of the HS Code being correct. We have evaluated our proposed model on the ground-truth data set provided by Dutch customs officers. Results show promising performance of 71% overall accuracy.**

***Index Terms*—machine learning, HS Code, text similarity, unsupervised, sentence embedding**

I. Introduction

The Harmonized Commodity Description and Coding System, also known as the Harmonized System of tariff nomenclature is an internationally standardized system of names and numbers to classify traded products. The system is developed by the World Customs Organization (WCO). It is a variety classification system of about 5,000 categories of six-digit products, organized in a hierarchical structure by department chapters (2 digits), headings (4 digits), subheadings (6 digits) and supported by implementing regulations and explanatory notes. It allows economic operators, customs officials and legislators from any country to identify the same product using a numeric code.

When importing goods into the European Union, a 10-digit code is required (HS codes supplemented with 4 digits). These codes are also called commodity codes. Based on the commodity code, customs knows how many import duties and other import taxes, such as import VAT, they must levy on the import of the product into the EU. Import duties differ per product (and therefore per commodity code). In addition, a commodity code indicates whether non-financial measures apply to the import and export of a product. Consider, for example, the obligation of export or import permits, health certificates or other documents. In addition to the value of the goods and the country of origin, the commodity code is one of the fields in a customs declaration that affects customs duties, so it is important that these fields are correctly filled in on a declaration. An import declaration must contain a commodity code (10 digits), however an eCommerce declaration is sufficient with the HS code (6 digits). Fraud in cross-border trade is one of the key priorities for the EU customs administrations due to the increasing risk of transnational crime and terrorism and the e-commerce-driven growth of customs declarations. Hence, EU customs administrations have to rapidly increase their capability to search for more accurate data sources to better assess these risks and increase their inspection hit rate. One of the challenges for customs administration is to detect whether the HS Code has been correctly declared.

To address this challenge, this paper seeks to accelerate the uptake of state-of-the-art data analytics and incorporation of new data sources for more effective and efficient European customs risk management. It provides tailored solutions, that

build on modern methods in machine learning and natural language processing, to help targeting officers and strategic analysts to identify potential risks by applying a semi-supervised algorithm to the unlabelled data sets.

The main contribution of this paper is to provide an automatic model that is able to identify the correctness of declared HS Code based on the goods description. The model is built upon the unsupervised mechanism that used a pre-trained embedding encoder with the unlabelled historical declarations, which can be beneficial to the generalisation. The proposed model is built on the Dutch customs data set, however, by changing the used historical training data set and the language specific embedding model, it can easily be adapted by other customs.

The rest of paper is organised as follows: The literature review is discussed in Section II. In Section III we describe the data sets that will be used in this paper, and explain our proposed method. Experiments and results are presented in Section IV, with conclusions and future work in Section V.

II. Related Work

With the growing volumes of trade and the unprecedented increase in cross-border eCommerce, customs administrations are facing major challenges of how to ensure the proper collections of duties and taxes, while at the same time not disturbing the trade flows. There is a growing interest in customs to explore the potential of data analytics to help addressing these challenges [1], [2]. HS Code has been widely used by customs administration to check the tax for the import declarations. For many years, researchers have explored the machine learning techniques to tackle the HS Code prediction task. Abdolshah et al. [3] presented a system classifying the shipping containers X-Ray images in order to investigate whether the imports have been correctly declared. Other work depends on the multimodal deep learning approach where both image and textual information have been taken into consideration. For example, Turhan et al. [4] developed an automatic HS code detection system based on textual analysis of the product's description together with its visual properties. Li et al. [5] leveraged a deep learning approach (convolutional neural network) to integrate textual features and image features to predict the correct HS Code. However, their approach was only validated on the four specific categories of products and can not be generalised. Similarly, Harsani et al. [6] proposed a method that is based on N-grams to identify the HS Code chapter "*64*".

The declaration data set provided by the Dutch customs administration only consists of textual descriptions and the declared HS Code. Also the majority of declarations have an extremely short text content, which only consists of a few words. Short text content, containing limited information, usually brings vagueness and uncertainty. Recent work on the prediction of HS Code correctness are only based on the textual information. Ding et al. [7] provided an automatic approach that is based on the background nets to classify the correct HS Code according the textual description. Shaalan [8] carried out comprehensive experiments to compare the various machine learning algorithms for the HS Code prediction. He used traditional textual feature representation - TF-IDF and demonstrated that the performance of support vector machine (SVM) is better than the other investigated approaches such as Na¨ıve Bayes, KNN, decision tree and random forest. However, the best result from this research only achieved a recall and F1 score of 51% and 66% respectively.

In addition, Spichakova et al. [9] proposed a method for the assessment of false HS Code to improve customs fraud detection process. Their model is based on the Doc2vec [10] where textual description will be represented by the vector, and similarity between the HS Code description and declared description will be calculated. Nevertheless, their approach depends on the HS Code nomenclature description (Section III-A) which is static and ignores the historical data. The HS Code system can match sentences with similar words but cannot correlate descriptions against categories with similar meaning. For example, the description of HS Code "8702" is *"Vehicles; public transport passenger type"*. When using this description to check the declaration such as *"Volvo"*, the system has difficulties identifying the correct HS Code since the free-textual declaration does not include any words related to "cars" or "vehicles". To avoid such issues, the language semantic should be captured. In this context, we propose a novel model that uses the historical declaration in this paper. To apply machine learning for the HS Code detection task, the textual description needs to be transformed to the numeri- cal vector representation prior to be fed into machine learning algorithms. The traditional approach of doing so is known as the bag-of-words method, where the text is represented by the occurrence of words in the text content. However, this approach ignores the language semantic information [11]. In recent years, increasing number of researchers have started using distributed features which have become more widespread with the availability of deep learning. Many researchers have used distributed features to tackle the NLP tasks, such as sentiment analysis [12], [13], machine translation [14], [15], text classification [16] and topic modelling [17]. In general, there are two levels of distributed representation used for textual content: word and sentence. Word distributed representation (also known as word embedding) maps an individual word to a vector, while sentence distributed representation (also known as sentence embedding) maps a chunk of text to a vector.

A. Word Embedding

Word embedding is now a trendy text feature representation. It has been widely used since Tomas Mikolov et al. [18] proposed Word2vec. This model is based on a three-layer neural network that uses the neighbouring words to predict the central word (known as CBOW) or vice versa (known as Skip-gram). However, it shows that the global language information cannot be captured. Glove [19], as an alternative approach, learns word representation by dimensionality reduction on the co-occurrence word counts matrix. As a result, the global statistic information from language has been preserved. Although

the word embedding is an effective approach to identifying similar descriptions, the customs declarations usually contains more than one word, such as *"METAL SHOCK ABSORBER PLATE", "WEDDING FLOWERS RED", "MOBILE PHONE ACCESSORIES"*. Word embedding is impractical at this circumstance and embedding for sentence level is required.

B. Sentence Embedding

The success of word embedding has motivated the generation of sentence embedding. Considering customs declarations, the descriptions usually contain a set of words, this paper pays more attention on the distributed representation for sentence level. The most simple way to generating sentence representation from word embedding is based on the aggregation approaches (e.g. concatenation, averaging, maximum). For example, a few studies [20], [21] show the competitive performance of word vector averaging in the task of text classification. Beyond the simple calculation, the development of pre-trained sentence embeddings [22]–[24] have existed only recently where the distributed feature model usually is trained on a general language corpus and can be used in a variety of downstream tasks such as calculating sentence similarity in this paper.

III. Methodology

In this section, we describe in detail the methods that we use in this work. We start with briefly explaining the data sets that are involved in our experiments. We then discuss the text feature representation (sentence embedding), followed by the explanation of our proposed approach to automatically examine the correctness of HS Code. Finally, we discuss the metrics used to evaluate the model.

A. Data sets

Four data sets are related to this work, the structure and examples for each data set are showing in Table I.

- **HS Code nomenclature** is a standardised numerical method that hierarchically categorises traded products as we described in Section I. It comprises approximately 5300 product descriptions and has been widely used by the customs authorities to identify the tax for the imports. The code has been separated as three hierarchical levels, the first 2 digits identify the chapter (e.g., 01 refers to animals), the next 2 digits identify the groups within the chapter (e.g., 0101 is the products related to horse) and overall 6 digits represent subsections within the group and give more specific categories. In this paper, we mainly focus on the first 6 digits although there are some extended digits that have been applied for the real-life scenario (e.g., Dutch customs used 10-digit commodity code in the declaration system).
- **Historical training data set** is provided by Dutch customs administration. In total, there are 149,044 historical commodity declarations, and each instance includes goods description and its declared HS Code. Based on observation, a small proportion of data is written in Dutch, while the majority of declarations have an English descriptions. We will use this data set to generate the vector representation which will be described in Section III-C.
- **Historical testing data set** is similar to the historical training data set except there are 257,251 declarations. There is no overlap observed between training and testing data set. Both, the historical training data set and the historical testing data set, are unlabelled. According to customs, the majority of the declared HS Code are assumed to be correct. The average length of description in our training and testing data sets are 5.79 words and 7.87 words respectively. As shown in Figure 1, the majority of goods description for both two data sets are under 20 words in length, indicating that users prefer to describe their products using brief messages.
- **Manually labelled data set**, Dutch Customs Administration also proposed a data set that consists of 100 declarations. Each instance was manually investigated and labelled by the experts as to whether the HS Code has been properly declared. The instance with incorrect HS Code was labelled as "y", otherwise it was labelled as "n". This data set was deliberately created as balanced distribution where each class has 50 declarations.

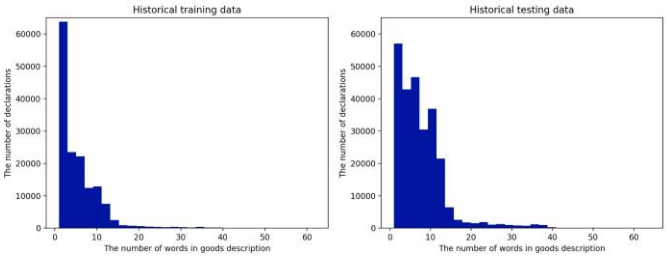

Fig. 1. Distribution of the number of words per description for two data sets

B. Sentence Embedding

When applying machine learning approaches to tackle the task of HS Code checking, the textual goods description must first be transformed into a compact representation of its content prior to its input to the downstream algorithm. Identifying an optimal feature representation is a critical step in text analysis. For decades, the dominant approach to feature representation for textual content has been based upon Vector Space Model (VSM) [25] where each text document is first tokenised (a process of segmentation) and then converted as a feature vector (e.g. bag-of-words or n-grams). However, simply adopting the VSM feature representation model would ignore the language information (e.g. semantic or syntactic) for the reason that it treats the features individually and largely ignores word order. Recent research [11] has explored the use of distributed representations where the text content is mapped into a vector by a pre-trained embedding model. The embedding model is trained on a general language corpus, which preserves intrinsic information. For example, word2vec [26] is one of the famous

TABLE I
FOUR DATA SETS INVOLVED IN OUR EXPERIMENTS

	Description	Code	Level
HS Code nomenclature	Animals; live	01	2
	Horses, asses, mules and hinnies; live	0101	4
	Horses; live, pure-bred breeding animals	010121	6
	.....	.....	.....
	Antiques; of an age exceeding one hundred years	970600	6
	Description	**HS Code**	
Historical training data set	WOODEN COFFEE TABLE WITH IRON STAND	940360	
	.....	.....	
	PLUS SIZE SUIT	620423	
	Description	**HS Code**	
Historical testing data set	COMMUNICATION EQUIPMENT	851712	
	.....	.....	
	MESSAGE CARD	391910	
	Description	**HS Code**	**Label**
Manually labelled data set	ESSENTIAL OIL SET	330290	y
	.....	.....	.....
	PENCIL SHARPENER	821410	n

models and widely used in NLP tasks. In our case, the goods description in declarations are made up by a few words, therefore we have used distributed representation on sentence level where each textual goods description will be transformed into a vector.

C. Approach

We assume that the similar import textual descriptions should be related to its corresponding HS Code group in the historical declarations. In this context, we propose an approach that is based on the similarity calculation to identify the correctness of HS Code.

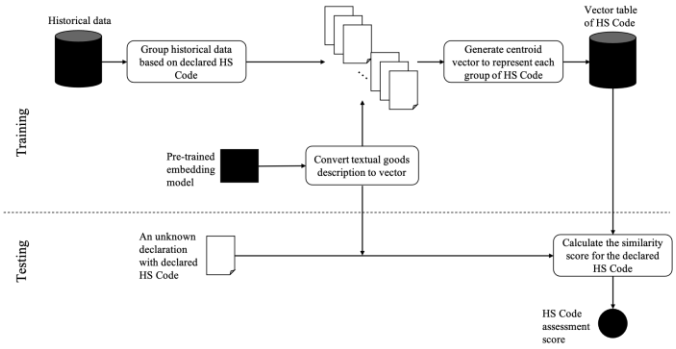

Fig. 2. The process of assessing the HS Code correctness

Figure 2 shows the process of how the first 2-digit HS Code correctness assessment has been performed. First 4-digit HS Code and overall 6-digit HS Code will follow the same process stream as the first 2 digits. As the training phase, the pre-trained embedding model is used to convert the textual commodity descriptions into the vector representations. We used historical records to generate the HS Code vector representation table. Prior to the explanation of our approach, we acknowledged that the historical data has errors (we are not using labeled records). However, we assumed that the declarations with incorrect HS Code only takes up a small portion of the data set and they have a minor impact on the total accuracy. The records in the historical training data set are separated to the various groups according to their declared HS Code. Then the declarations with the same declared HS Code will be under the same group (HS Code). Each HS Code group may have a different size due to the imbalanced distribution in the historical data set. We analysed the distribution of HS Code groups in the historical training data set. Figure 3 shows that approximately 15% of declarations are declared as electrical equipment (HS Code: 85) while only less than 4% of the declarations are declared as printed books (HS Code: 49). Every single declared description will be converted to the vector representation by the use of pre-trained language embedding model. The vectors within the same HS Code group will be used to calculate the mean values and will be referred to as the ”centroid vector” (Equation 1) where V_i is the vector of goods description, and N is the number of vectors within the group. Once the training process is completed, the HS Code with its associated vector will be generated as a lookup table where each HS Code will be represented by the fixed dimensional vector.

$$Centroid\,V = \sum_{i=1}^{N} \frac{V_i}{N} \tag{1}$$

In the testing process, a goods description will be converted to the vector using the same pre-trained embedding model as the training phase, then the vector will be used to calculate the similarity scores against each HS Code vector in the lookup table. HS Code will be ranked based on the similarity score with the order from highest to lowest. At last, the associated similarity score to the declared HS Code will be identified and used as the correctness score. Figure 4 illustrates the process of correctness calculation.

The aim of our research is to predict the degree of correctness based on a textual description associated with its HS Code. However, simply calculating the cosine similarity as degree of correctness is not applicable due to that the

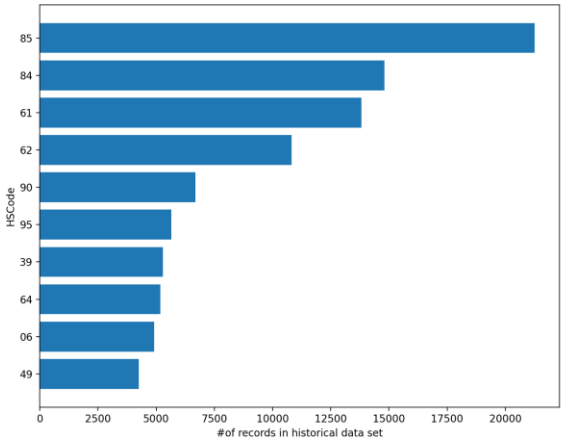

Fig. 3. The most 10 common HS Code in the historical data set

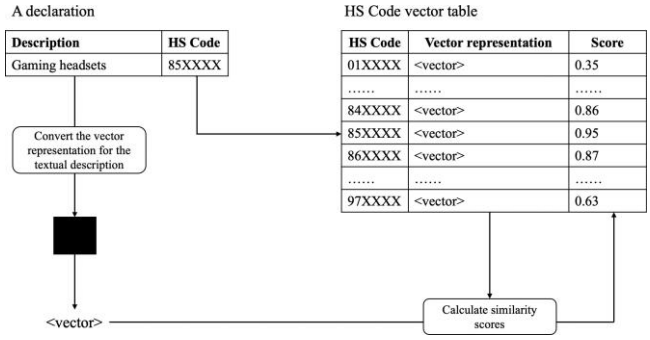

Fig. 4. The mock-up process of how HS Code correctness score has been computed in the testing phase

negative value would appear in the results. The final score preferably has to be ranged into [0-1]. To tackle this problem, we have used the angular similarity (Equation 2) instead of cosine similarity, which is suggested by Cer et al. [22]. In addition, we used Min-Max normalised approach to scale the similarity score. The final normalised angular similarity score is distributed equally in the range from 0 to 1 which can be used as correctness score. The higher value (close to 1) means the HS Code is highly likely to be correct, otherwise the HS Code is more likely to be incorrect.

$$angular\,similarity = 1 - \frac{cos^{-1}(cosine\ similarity)}{\pi} \quad (2)$$

D. Evaluation Metric

Given a specific goods description and its HS Code, our proposed model predicts a score to indicate the degree of correctness for the declared HS Code. However, there is no ground-truth labelled data to be used for evaluation. Our test data only include the binary label to indicate whether the HS Code is correct or not. To transform the continuous score prediction into the binary prediction, we used the various thresholds as the boundaries to convert the regression problem into the classification problem. The results of the experiments are reported using the standard text classification measures: accuracy, precision, recall and F1-score. In addition, we used the receiver operating characteristic curve (ROC curve) to investigate the performance of our model with various thresholds.

IV. Experiments and Results

This section will elaborate the experiments and explain the results. In order to assess the correctness of HS Code in declarations, we proposed an approach based on calculating the similarity score between the descriptions. We carried out three assessment experiments in this section to evaluate the approach. Firstly, we examined the quality of HS Code vector representation itself. Secondly, we used an unlabelled data set to demonstrate that the predicted correctness score can be used as an indicator to show the degree of HS Code being correct. Thirdly, we evaluate the approach as classification task by the use of the labelled data.

Prior to the assessment experiments, we have acknowledged that the selection of pre-trained embedding model to represent the goods description for customs import declaration is a key factor for the model performance. We have carried out some preliminary experiments, and the results show the best performance is achieved by the use of Universal Sentence Encoder (USE) [22]. Although Bert [23] is a well-known model and has been widely used in the natural language processing domain, it's showing a lower performance compared to the USE, particularly in the task of short text mining. In addition, some scholars [27], [28] suggest that the embedding model pre-trained on the similar source body of text as the downstream tasks will increase the performance. However, building our own embedding model would require a much larger language corpus in the customs declaration domain than currently available. Therefore, we propose this as one of our future plans.

A. Experiment 1 - HS Code vector representation

As described in Section III-C, we proposed a model that is based on the historical data set to generate the HS Code vector representation table which will be used to calculate correctness score. To evaluate the quality of HS Code vector representation table itself, we assume that the chapter description (the first 2 digit) will be semantically close to its subsections descriptions (4 digit or 6 digit). According to our observation, we identified that most of the HS Codes satisfy this assumption. As illustrated in Figure 5 (left), we graphed the most 10 similar HS Codes for the chapter ("64"), these similar items are all coming from its sub-categories such as "6403", "640399", "640291". In addition, the subsections within the same chapter also have the high similarity score, for example, the similarity score between "640391" and "640399" is 0.85. However, this observation is not always consistent. There are a few HS Codes that have unmatched neighbouring HS Codes, An example is showing in Figure 5 (right). There are some other categories

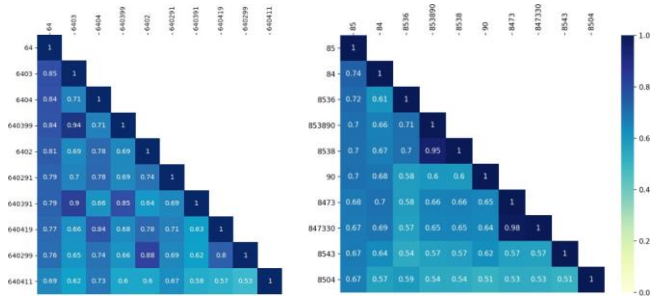

Fig. 5. HS Code (64) similarity scores using embeddings from the universal sentence encoder

(e.g., “84”, “90”) that are appeared in the most similar items for the HS code “85” rather than its subsections such as ”8543” or ”8504”. This discrepancy will jeopardise the correctness of our prediction because the closed chapter code will skew the similarity score while the subsection is the correct prediction.

B. Experiment 2 - unlabelled data set evaluation

The second evaluation experiment is conducted on the unlabelled testing data set. We predicted the correctness of a HS Code on three different digits levels. In the 2-digit prediction, the first 2 digits of HS Code will be checked, and 4-digit prediction will check the correctness of the first 4 digits. 6-digit prediction will check the entire HS Code. In this case, three HS Code vector tables with different levels will be developed once we completed the aforementioned training process. In each table, the HS Code will be represented by the vector.

Although the data is unlabelled and not manually investigated, the majority of historical declarations are suggested by Dutch Customs Administration to be properly declared with the correct HS Code. To verify this assumption, Figure 6 shows the distributions of correctness scores across three levels of HS Code (2-digit, 4-digit, and 6-digit) prediction. It is expected that most of the declarations have high correctness score (above 0.8), which proves that our proposed correctness score based on the similarity calculation can be treated as an indicator to show the degree of the HS Code being correctly declared.

C. Experiment 3 - labelled data set evaluation

Another evaluation was carried out on the manually labelled data set where every declaration in this data set has been checked by customs targeting officers and the HS Code was tagged as correct or incorrect by means of physical inspection. To evaluate our approach on the binary classification problem, we decided to use the ”threshold” to predict whether the HS Code is correct based on the correctness score. For example, if the threshold is ”0.8”, the declarations with the correctness score higher than 0.8 will be classified as ”correct”, otherwise it will be classified as ”incorrect”. In this context, we examined our approach across three levels of digits prediction. The overall comparable results are showing in Figure 7. All three

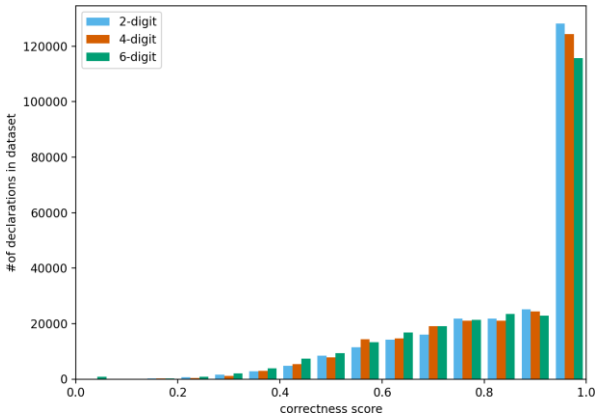

Fig. 6. The distribution of correctness score in the historical testing data set

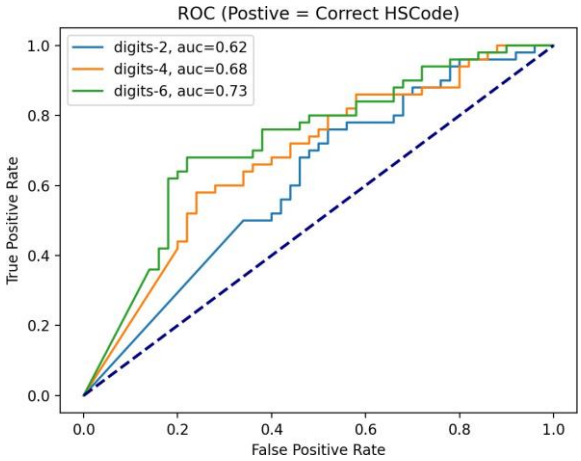

Fig. 7. The ROC results of three digits classification across varying thresholds

levels of digits predictions achieved better performance rather than randomly classification. In addition, the best result is achieved by the use of 6-digits prediction where the AUC is 0.73.

The results are reported in Table II, the threshold is created from 0 to 1 in intervals of 0.1. For each threshold, we conducted classification experiments on three levels of digit. Overall the more digit prediction outperforms the less digits, specifically, the best results 0.71 accuracy is achieved by the use of 0.8 as threshold and 6-digit prediction.

It is interesting notice that the best results are achieved by 6-digit prediction, which is counter-intuitive as predicting the chapter (2 digits) is expected to be easier than predicting the subsection (6 digits). We suggest that the reason why the prediction of 6 digits outperforms the prediction of 2 digits is because the vector representation for 2-digit HS Code is generated from more general descriptions while the 6-digit HS Code is generated from more specific descriptions. Therefore,

TABLE II
RESULTS OF BINARY CLASSIFICATION ACROSS 3 LEVELS VARYING DIFFERENT THRESHOLDS

Threshold	2-digit				4-digit				6-digit			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
0	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33
0.1	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33
0.2	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33	0.5	0.25	0.5	0.33
0.3	0.5	0.25	0.5	0.33	0.52	0.76	0.52	0.38	0.53	0.76	0.53	0.4
0.4	0.51	0.75	0.51	0.35	0.53	0.76	0.53	0.4	0.55	0.76	0.55	0.44
0.5	0.52	0.59	0.52	0.4	0.55	0.76	0.55	0.44	0.58	0.69	0.58	0.51
0.6	0.57	0.68	0.57	0.49	0.57	0.65	0.57	0.5	0.59	0.65	0.59	0.55
0.7	0.57	0.61	0.57	0.52	0.63	0.65	0.63	0.61	0.66	0.67	0.66	0.66
0.8	0.6	0.61	0.6	0.59	0.64	0.64	0.64	0.64	0.71	0.71	0.71	0.71
0.9	0.56	0.56	0.56	0.56	0.67	0.68	0.67	0.67	0.64	0.66	0.64	0.63
1	0.58	0.58	0.58	0.58	0.61	0.63	0.61	0.6	0.61	0.65	0.61	0.58

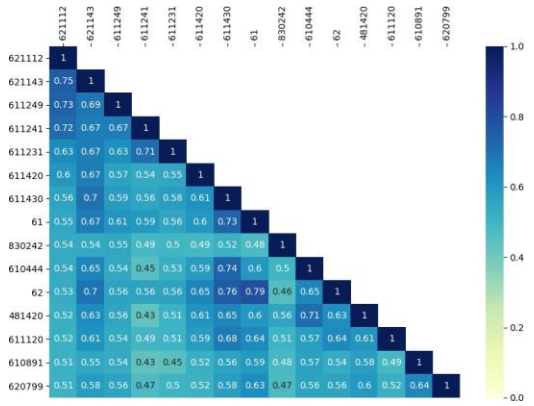

Fig. 8. The most similar HS Code to ”621112” based on the similarity scores using embeddings from the universal sentence encoder

the later is able to distinguish more detailed descriptions. In addition, as we described before, we observed that the HS Code itself can be skewed by the neighbour chapters. For example, Figure 8 shows the top 15 similar HS Codes for the ”621112”. The centroid of chapter ”621112” has shifted due to the goods descriptions of the 2-digit goods HS Code 61. When we predict the chapter level (2-digit), the expected outcome will be chapter 61 instead of 62 because the similarity score is 0.55 for chapter 61 and only 0.53 for chapter 62. We assume that the improvement of historical data quality, and enhancement of the pre-trained embedding model will alleviate this discrepancy issue, we will research on these two directions in our future work.

V. CONCLUSION AND FUTURE WORK

The purpose of this paper was to propose a novel solution based on machine learning to automatically identify the correctness of HS Code in the customs declarations. The conceptual idea of this model is based on that the textual descriptions should be related to its corresponding HS Code group in the historical declarations. In this paper, the off-the-shelf sentence embedding (universal sentence encoder) was implemented to convert the textual data into numerical features and historical declarations were used to generate the HS Code representations (as we named HS Code centroid vector). Given a textual goods description with its declared HS Code, our proposed model will compute the similarity score between the description across every single HS Code centroids and predict a reliable score to indicate what degree the HS Code has been declared correctly.

We highlight the following observations: Firstly, using the normalised angular similarity score is an applicable approach to indicate the degree of correctness for HS Code. Secondly, language in descriptions of historical declarations is not always correct as there are some discrepancies. For example, the 2-digit HS Code is semantically surrounded by other 2-digit HS Code rather than its subsections (e.g., 4-digit or 6-digit HS Code). We assumed these discrepancies are caused by the HS Code nomenclature system where the description is ambiguous and impacts the error declarations in historical data. Thirdly, we identified that 6-digit HS Code achieved a better performance compared to the same data set where only the first 2 digits are being predicted of the HS Codes.

Our future work in this area is focused on three-folds: (1) improving the quality of historical data set. At moment, the current data set includes some declarations that are written by Dutch instead of English. Although the volume of those declarations is small, it will impact the accuracy of HS Code vector representation. (2) enhancing the sentence embedding model, using the domain specific data to fine tune the current model is a promising way to increase the prediction accuracy. (3) exploiting human-in-the-loop learning such as active learning to provide a continual learning capability in this domain.

ACKNOWLEDGMENT

This research was partially funded by the PROFILE Project (nr. 786748), which is funded by the European Union’s Horizon 2020 research and innovation program. Ideas and opinions expressed by the authors do not necessarily represent those of all partners.

REFERENCES

[1] F. Heijmann, Y. Tan, B. Rukanova, and A. Veenstra, “The changing role of customs: Customs aligning with supply chain and information management,” *World Customs Journal*, vol. 14, no. 2, 2020.

[2] B. Rukanova, Y.-H. Tan, M. Slegt, M. Molenhuis, B. van Rijnsoever, J. Migeotte, M. L. Labare, K. Plecko, B. Caglayan, G. Shorten *et al.*, "Identifying the value of data analytics in the context of government supervision: Insights from the customs domain," *Government Information Quarterly*, vol. 38, no. 1, p. 101496, 2021.

[3] M. Abdolshah, M. Teimouri, and R. Rahmani, "Classification of x-ray images of shipping containers," *Expert Systems with Applications*, vol. 77, pp. 57–65, 2017.

[4] B. Turhan, G. B. Akar, C. Turhan, and C. Yukse, "Visual and textual feature fusion for automatic customs tariff classification," in *2015 IEEE International Conference on Information Reuse and Integration*. IEEE, 2015, pp. 76–81.

[5] G. Li and N. Li, "Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network," *Electronic Commerce Research*, vol. 19, no. 4, pp. 779–800, 2019.

[6] P. Harsani, A. Suhendra, L. Wulandari, and W. C. Wibowo, "A study using machine learning with n-gram model in harmonized system."

[7] L. Ding, Z. Fan, and D. Chen, "Auto-categorization of hs code using background net approach," *Procedia Computer Science*, vol. 60, pp. 1462–1471, 2015.

[8] F. Altaheri and K. Shaalan, "Exploring machine learning models to predict harmonized system code," in *European, Mediterranean, and Middle Eastern Conference on Information Systems*. Springer, 2019, pp. 291–303.

[9] ——, "Exploring machine learning models to predict harmonized system code," in *European, Mediterranean, and Middle Eastern Conference on Information Systems*. Springer, 2019, pp. 291–303.

[10] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[11] H. Chen, S. McKeever, and S. J. Delany, "The use of deep learning distributed representations in the identification of abusive text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 125–133.

[12] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, "Learning user and product distributed representations using a sequence model for sentiment analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 34–44, 2016.

[13] Z. Cui, X. Shi, and Y. Chen, "Sentiment analysis via integrating distributed representations of variable-length word sequence," *Neurocomputing*, vol. 187, pp. 126–132, 2016.

[14] R. Wang, A. Finch, M. Utiyama, and E. Sumita, "Sentence embedding for neural machine translation domain adaptation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 560–566.

[15] C. Shi, S. Liu, S. Ren, S. Feng, M. Li, M. Zhou, X. Sun, and H. Wang, "Knowledge-based semantic embedding for machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2245–2254.

[16] L. Ge and T.-S. Moh, "Improving text classification with word embedding," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1796–1805.

[17] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Using word embedding to evaluate the coherence of topics from twitter data," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 1057–1060.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Charagram: Embedding words and sentences via character n-grams," *arXiv preprint arXiv:1607.02789*, 2016.

[21] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016.

[22] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder for english," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 169–174.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "Kepler: A unified model for knowledge embedding and pre-trained language representation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.

[25] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[27] H. Chen, S. McKeever, and S. J. Delany, "A comparison of classical versus deep learning techniques for abusive content detection on social media sites," in *International Conference on Social Informatics*. Springer, 2018, pp. 117–133.

[28] S. Lai, K. Liu, S. He, and J. Zhao, "How to generate a good word embedding," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 5–14, 2016.