Exploring Machine Learning Models to Predict Harmonized System Code

Chapter *in* Lecture Notes in Business Information Processing · April 2020

DOI: 10.1007/978-3-030-44322-1_22

Khaled Shaalan
British University in Dubai

Exploring Machine Learning Models to Predict Harmonized System Code

Fatma Altaheri(✉) and Khaled Shaalan

British University of Dubai, Dubai, UAE
20181351@student.buid.ac.ae,
khaled.shaalan@buid.ac.ae

Abstract. The Harmonized System (HS) Code is widely used across all customs administrations because of the several benefits including a more convenient and easier approach for calculating duties as well as preventing the potential loss of revenue. This paper aims to explore various machine learning models to predict the HS Code based on the customers' input commodity descriptions. This prediction model helps in reducing the complexity, gaps and many other challenges in using HS Code in any Customs administration. This study follows the Cross-Industry Process for Data Mining methodology which comprises six phases, namely business understanding, data understanding, data preparation, building prediction model, performance evaluation and model deployment. The results of the study indicate that machine learning models are effective tools in predicting HS Code based on user's inputs. The linear support vector machine model was able to achieve the highest accuracy of 76.3%.

Keywords: Harmonized System Code (HS Code) · Customs, revenue protection · Machine learning · Predictive models · Trade

1 Introduction

The emergence of globalization has paved way to opening a number of opportunities towards achieving economic growth and prosperity. However, there are still certain challenges that pose as barriers to the effective implementation of processes within local governments. Customs is one of the most important government agencies, in the global perspective, that governs international trade and services process such as the declaration of goods and services, particularly in the aspect of trade and commerce. Generally, the verifying commodity descriptions ensures that goods and/or services are in compliance with government regulations to prevent improper or unlawful entry to the country of destination (Che et al. 2018). In this sense, Customs has the responsibility of ensuring that declared goods and/or services to be imported or exported are classified accordingly based on commodity descriptions.

Goods classification is one of the most important obligation of importer and exporter compliance. As such, it has become vital to customs departments worldwide to ensure that there is a delicate balance of reducing classification uncertainties and promoting effective classification systems. In response to this, several customs departments across the globe has integrated the use of modern technology by adopting

M. Themistocleous and M. Papadaki (Eds.): EMCIS 2019, LNBIP 381, pp. 291–303, 2020.
https://doi.org/10.1007/978-3-030-44322-1_22

the Harmonized System (HS) Code. According to Ding, Fan and Cheng (2015), the HS Code, also known as the Harmonized Commodity Description and Coding System, was developed by Brussels-based World Customs Organization (WCO) in order to cope with the rapidly increasing international trade worldwide. In line with this, Weerth (2008) explained that goods in a customs tariff must be described fully within the customs declaration so they can be classified accordingly. For example, a wooden chair can be classified according to material condition or its function as furniture (Weerth 2008). This means that the use of HS Code has become important for customs departments because it promotes easier ways of calculating duties, fees and taxes, determining appropriate permits, licenses and certificates required and collecting trade statistics (Ding et al. 2015).

Traditionally, declared goods are analyzed and inspected by inspectors using inspection images wherein results are used as grounds for decision making (Che et al. 2018). Using this approach, a number of weaknesses and concerns can be identified such as difficulty in identifying hazards in substances and ineffective classification of declared goods due to possible inefficiencies in the designation of harmonized system codes (Che et al. 2018; United Nations 2013). Aside from these weaknesses, there are also some challenges in the application of HS Code in relation to achieving satisfactory accuracy. These challenges include HS complexity, gaps in terminology and the evolving nature of the HS Code among others (Ding et al. 2015).

This study aims to contribute to reducing HS complexity and gaps by exploring various machine learning-based prediction models. In order to provide a better outlook as to the impact of the adoption of the machine learning-based HS Code Prediction model, the case of Dubai Customs is used. This means that this study employs a case study approach focusing on the Customs Department in Dubai, UAE. It is expected that the integration of technological advancement to the HS Code classification can contribute to addressing HS Code complexity and to enhancing its accuracy.

The paper is divided into 5 sections. Section 1 provides introductions about the HS Code and the importance of this study. Section 2 explores the related work that have been done to address the HS code research gaps. Section 3 illustrates the research methodology in details and the various machine learning models that have been implemented. Section 4 discusses the main findings. The last section provides the conclusion of this work and the future prospects.

2 Related Work

Machine learning methods are currently being integrated in various programs and systems in order to help organizations make better decisions particularly with regards to predictive analysis and pattern recognition. For example, machine learning methods are majorly being integrated in the security field such as to enhance facial recognition from using large amount of data from various sources which is difficult for humans to do manually (Mohammed et al. 2017). In addition, machine learning is also being integrated to various systems to promote automation and improve efficiency and accuracy. For example, machine learning techniques are being integrated in HS codes to foster accuracy, intelligence and automation (Zang et al. 2008). This suggests that multiple machine learning strategies integrated in various programs and systems can contribute to enhancing system performance, continuous learning and promoting more informed decision making.

In the perspective of customs organizations/departments, the integration of machine learning has become vital in order for them to stay competitive in the international trade facilitation (Ding et al. 2015). As explained by Ding and colleagues, different machine learning techniques are already being applied in capturing and learning long term and stable criteria for text categorization. For example, the use of keywords is one of the most common approaches of machine learning through the adoption of the vector space model (VSM) (Ding et al. 2015). In addition, machine learning approaches are also being integrated in HS in pursuit of addressing HS code prediction problems (Luppes 2019).

Furthermore, machine learning strategies are also being integrated in HS in order to enhance automation process. According to KPMG International (2018), machine learning can be used to develop a knowledge base towards learning and developing a set of algorithms from large amount of data in order to make informed predictions. KPMG International (2018, p. 37) explained that "A combination of natural language processing and machine learning makes it possible to automate the capture, array and analysis of unstructured data and transform it into structured data that may be used in a tax application". This suggests that the integration of machine learning in HS can contribute to process efficiency by means of improving quality, consistency and accuracy of code classification due to reduced likelihood of human errors. There are several machine learning tools and techniques that have been developed. Yet, machine learning can be viewed as the one that provides the technical basis for data mining (Witten et al. 2011).

In order to promote effective application and integration of machine learning to HS, it is important to understand how it works. In the general perspective, machine learning can be viewed as the new technology for mining knowledge from data in order to learn and generate accurate predictive outcomes (Witten et al. 2011). As such, machine learning can be viewed as being directly correlated with data mining. In line with this, machine learning techniques integrated with HS can also foster smart matching and classification through building data driven smart customs (Youyi 2017). As explained by Youyi (2017), a data driven smart customs can promote intelligent law enforcement, intelligent risk control and intelligent revenue collection wherein data collections can lead to intelligent applications (see Fig. 1).

Therefore, data mining is an important part of creating a machine learning framework that can contribute to promoting accurate and efficient classification and detection of tariff code as integrated in HS code systems. Collecting large amount of data set and generating useful information from collected data will not become a challenge due to the integration of machine learning in HS. Using machine learning based approaches and models can help enhance predictive analysis on streaming data and detect changes and inconsistencies in data to address HS code classification and description and business intelligence problems.

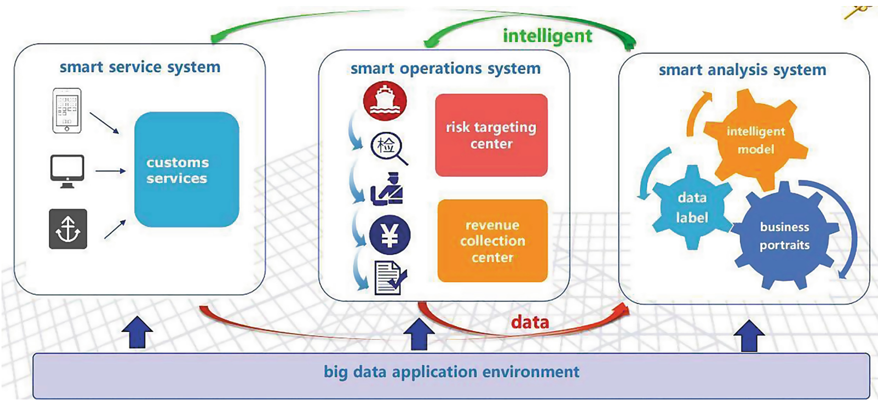

Fig. 1. Architecture of data driven smart customs (Youyi 2017)

3 Research Methodology and Experiments

This study aims to investigate the possibility of implementing machine learning model to predict the HS code for the commodities based on the provided descriptions by the users. To perform this text mining task, we adopted the Cross-Industry Process for Data Mining methodology (CRISP-DM) (Shearer 2000). As illustrated in Fig. 2, this methodology consists of the following processes, namely business understanding, data understanding, data preparation, data modelling, performance evaluation and deployment. Business understanding refers mainly to the importance of the problem that is being address which is reducing the loss revenue by building machine learning model to correctly predict the HS Code. Data Understanding refers to the process of analyzing the available input information in order to produce an efficient machine learning model. Data preparation represents the preprocessing steps which are typically performed in order to remove any factor which may degrade the performance of the machine learning model. The machine learning modelling, evaluation and deployment involves the selection the appropriate machine learning techniques in order to investigate and evaluate these models and figure out the best machine learning model that produces the highest performance for deployment.

This section starts by describing the source of the HS Codes dataset that is used to conduct the machine learning experiments. Furthermore, a detailed analysis on how the obtained dataset can be used to answer the driving research question. And this is followed by the data preparation tasks that includes data cleansing, sample processing and tokenization that will be used to build the inputs features to the machine learning models. This section also explains in details the machine learning models that are adopted and implemented throughout the HS Code prediction models.

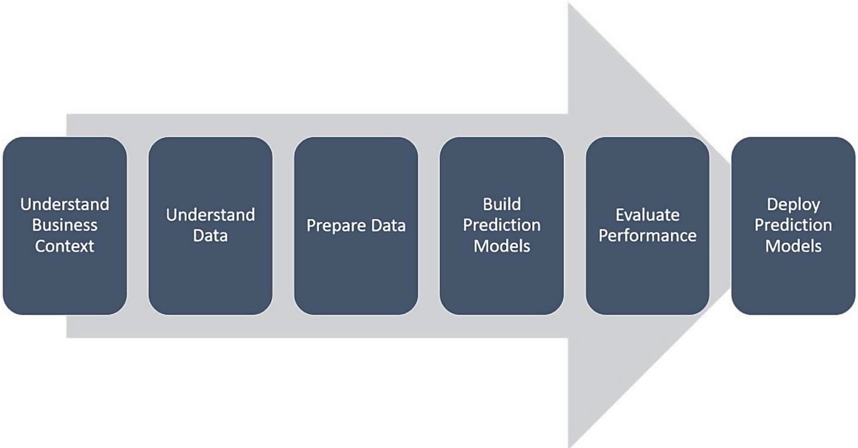

Fig. 2. Research methodology

3.1 Business Understanding

The importance of this work scope is to reduce the loss revenue caused by misclassifying the goods. This misclassification results in loss of duties. Such misclassification normally occurs to either the ability to the user not to correctly determine the required HS Code or in another situation the manipulation by the users to reduce the amount of duty that he/she is required to pay. In this work, the focus is on the first issue since the second issue requires other authentication techniques to determine whether the user is intentionally misclassifying the goods.

3.2 Data Understanding

The data used in this work are provided by Dubai Customs through the Artificial Intelligence (AI) hackathon competition that was conducted on October 2019. This data consists of 22,346,194 records where each record has two attributes; the Harmonized System Code (HS Code) and the user inputs description. In this section, we will describe the processing techniques that we adopted in order to review, analyse and prepare the input data.

3.3 Data Analysis

During the analysis of the provided data, we have noticed several major factors which must be taken into consideration during the development of the machine learning based model. These factors are mainly related to the quality of the user's description and the number of expected classes (labels).

About the user's description, several issues have been noted. For instance, some of the descriptions do not contain any valuable keywords from the English dictionary. In another words, the spelling mistakes in these descriptions are too severe to the point that they could not be understood by the human mind. Also, we have noticed that the same description has been used several times to describe totally different items.

Regarding the expected number of classes, the provided data has almost 8,000 labels. And this underlines the problem associated with establishing any machine learning model to predict the label based on the provided description. Additionally, it has been noticed that they are severe variations in terms of the number of records that are associated with each class. For instance, some of the classes have around 10 records while other classes may have up to 700 records. This factor must be addressed in order to avoid having any bias behavior in the machine learning model.

3.4 Data Cleansing

The above-mentioned factors have been taken into consideration during the cleaning of the data. And in this section, we will present the procedural steps that have been performed during the cleaning of the data. These steps can be summarized as follows:

(1) remove duplications (2) remove punctuation and remove stop words (3) remove non-English words (4) remove numbers and (5) lemmatization.

Remove Duplications: Duplicated records might have negative impact on the overall performance of the machine learning models. Therefore, we have removed all duplicated records based on the users' description as a pre-processing step.

Remove Punctuation and Stop Words: Eventually, the punctuation and stop words have no impact in the training behavior of any machine learning model since they are common words that do not add any distinct features to the description. Therefore, the punctuation and stop words have been removed in order to speed up the training process.

Remove Non-English Words: Normally, the descriptions are provided by users who have basic understanding of the English language. And this may result in writing words with spelling mistakes which may confuse the behavior of the training because at the training stage, these non-English words will be interpreted with different words that have different meaning. Therefore, these non-English words were removed during the cleansing step.

Remove Numbers: In this domain, it has been noticed that numbers do not provide any valuable features to the goods. Thus, the number have been removed as well in order to speed up the training for the machine learning model.

Lemmatization: This process is used to return the word to its origin. Thus, applying lemmatization is expected to reduce the vocabulary size and improve the performance since words that have the same origin must be treated equally.

In addition to the above-mentioned cleansing steps, all text has been converted to lower case in order to reduce the vocabulary size since any machine learning model is case sensitive.

3.5 Sample Processing

Having classes with very big variation in terms of records is expected to create a significant impact on the training behavior. For instance, if we assume using the

K- nearest neighbor to predict a binary class label where the first class has 100 records and the second class has only 10 records. In this case, the model may mis-classify the data towards labeling the reading by the 100-records label class. Accordingly, this sampling issue must be addressed to ensure consistency in terms of the classification behavior. To address this problem, we have reported the minimum and maximum records for the 8,000 classes that are provided in the dataset. During this step, it has been noticed the gap between the minimum and the maximum reaches to 750 records. Thus, we have performed down sampling to achieve 100 records as a gap between the minimum and maximum class size.

In this work, due to computational limitation of the used machine, we have selected only 500,000 at random as the input data. Additionally, once all pre-processing steps are performed, 217,700 records are kept as the final data set input.

3.6 Tokenization

In NLP, tokenization is an important step in order to determine weight (importance) of each word in the text. In general, two major techniques have been widely used to perform tokenization. These techniques are: (1) bag of words and (2) Term Frequency - Invert Document Frequency (TF-IDF). The bag of words is simply a counting method where each word is given a value that represents the frequency in which the word appears in the text. Accordingly, the use of this technique depends whether the number of appearances of each word in the text is considered as distinct features in each model (Alqaryouti et al. 2019).

However, the TF-IDF is considered as a more advanced technique since it uses the frequency of the word in order to determine the uniqueness of each word in the text. The TF-IDF starts by calculating the frequency of each word in the text ($Fr(w)$). This is calculated by dividing the number of times word (w) appears in the text over the total number of words. This denotes the term frequency in the calculation (TF). The inverse document frequency (IDF) is calculated as follows:

$$IDF(w) = \log(\frac{N}{Nw})$$

Where N denotes to the total number of records (the users' HS Code descriptions), and Nw denotes the number of records that has the word w. This inverse is normally calculated to associate a weight value with each word where words with less appearance have more weight. Eventually, the TF-IDF score for a word w can be represented as follows:

$$T\mathrm{F} - IDF(\mathrm{w}) = TF(\mathrm{w}) \times IDF(w)$$

The equation shows that the TF-IDF aims to magnifies the weight of the words that appears in a smaller number of records (Alqaryouti et al. 2018).

3.7 Experiments

In this section, we present and discuss the machine learning models that are used in this work and the corresponding performance of each model. These experiments were performed using Python 3.5 with scikit learn library for model specifications. To evaluate the performance for each model, we have adopted the following evaluation metrics: precision, recall, F1-measure and accuracy. The machine learning prediction approach is illustrated in Fig. 3.

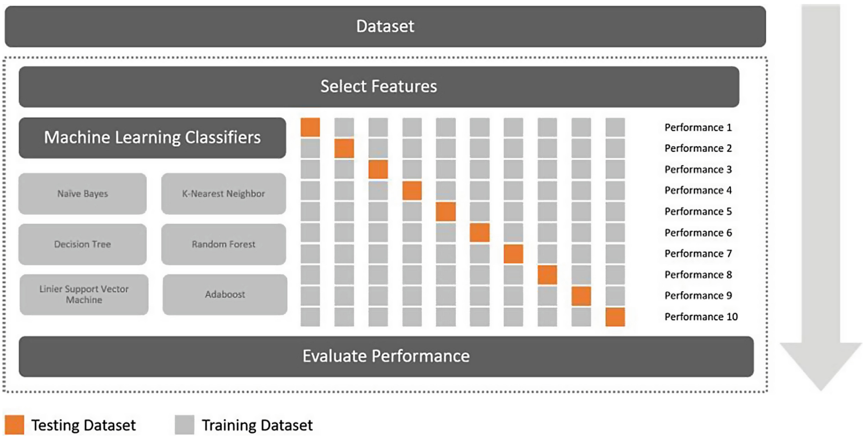

Fig. 3. Machine learning prediction and evaluation approach

3.7.1 Performance Measures

For each HS Code label, precision refers to the number of correctly classified description that belongs to this label. For instance, the precision for HS Code (00000000) refers to the percentage of correctly classified goods that belong to this label where the equation of calculating the precision is show below:

$$P = \frac{TP}{TP + FP}$$

Where, number of goods that are classified as (00000000) label is referred as True-Positive (TP). The False-Positive (FP) denotes the number of goods (descriptions) that have been misclassified as (00000000) label.

Recall (R) denotes the correctly classified labels. For instance, the recall for the class (00000000) in the above example refers to the correctly classified goods that belong to this label and is calculated as per the below equation:

$$R = \frac{TP}{TP + FN}$$

Where, false-negative (FN) denotes the number of goods (records) that belong the class (00000000) and are misclassified as other classes.

The F1-score is used a measure to determine the quality of the classification, where a higher F1-measure score highlights better classification quality. The F1-measure score is calculated according to the following equation:

$$F1 = 2 \times \frac{R \times P}{R + P}$$

The accuracy refers to the number correctly classified goods by the machine learning model and it is calculated as follows:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

3.7.2 HS Code Prediction Models

In the experiments that are presented in this work, we have performed 10-fold cross validation technique. This technique which is also denoted as rotation estimation is typically used to assess the performance of the machine learning model by generating different independent equal sets (k = 10). In general, any machine learning model works by using the training set to learn the behavior of the data. Whereas, the test set is the unseen data that is used to measure the accuracy of the learning behavior of the machine learning model. In cross validation, the experiment is performed k times where in each round the data is divided into training and testing sets. In each round the selected testing set must be different where the performance is monitored over the k runs in order to better analyze the model performance.

The main ingredients of this work are to use the following machine learning models in order to evaluate and answer the proposed research question. These machine learning models are: Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, Linier Support Vector Machine and Adaboost. In these machine learning models, we will perform the prediction model on two experiment settings. The first setting, we will test whether the machine learning model is able to predict the entire HS Code Correctly. Whereas, in the second setting, we will test the ability of only predicting the header of the HS Code. The header denotes the first four digits of the HS Code which represent the chapter (2 digits) and section (2 digits). The chapter and section represent the commodity type (e.g. coffee, skin products).

The input data for the machine learning models are expressed as a tuple that consists of the following features: <HS Code, Description Text> where, once the tokenization is performed, the description text is converted to several features that represent the weights of the words in each description. HS Code represents the label that will be predicted.

In its core, the Naiive Bayes classifier uses the Bayes theorem to predict the probability of membership. The main idea in this classifier is to use conditional probability in which the probability is calculated between each input feature and the corresponding classes. In other words, if we assume that we have 5 features and 2 classes, the mechanism of this classifier works by determining the conditional probability of each feature and the 2 classes. This classifier is very efficient in text- based modelling where the input is expected to have distinct words that highlight the advantages of using the conditional probability concept. To test the performance of the Naïve Bayes classifier in predicting the HS Code, we ran the experiment where Table 1 shows the performance metrics. From the table, we can see that the accuracy of this classifier is relatively low (52.43%) and this highlights the fact that the input description consists of relatively common words that do not distinguish between the input record efficiently.

Table 1. Performance evaluation adopted machine learning models

Machine learning model	Experiment settings	Precision	Recall	F1-Measure	Accuracy
Naïve Bayes	HS Code header	73.66%	55.66%	63.40%	66.21%
	Entire HS Code	59.67%	29.45%	39.44%	52.43%
K-Nearest Neighbor	HS Code header	72.83%	57.75%	64.42%	71.72%
	Entire HS Code	55.30%	26.66%	35.97%	57.94%
Decision Tree	HS Code header	70.21%	46.92%	56.25%	61.62%
	Entire HS Code	51.00%	20.72%	29.47%	47.84%
Random Forest	HS Code header	96.35%	64.39%	77.19%	79.99%
	Entire HS Code	94.00%	38.19%	54.31%	66.21%
Linear Support Vector Machine	HS Code header	**96.35%**	**70.55%**	**81.46%**	**84.58%**
	Entire HS Code	**95.06%**	**51.41%**	**66.73%**	**75.40%**
Adaboost	HS Code header	73.66%	67.44%	70.41%	75.40%
	Entire HS Code	51.00%	25.10%	33.65%	57.02%

Table 1 shows the performance measure of this classifier where the label of each class only represents the header. Recall that the header represents the first 4 digits of the HS Code. And this header classifies an abstract level the goods type. From this table, we can see that the by considering the header of the label, the performance of the classifier has improved significantly, where the classifier achieves 66.21% accuracy. This is related to the fact that by aggregating the records and only perform the classification on the HS Code header, we reduce the pressure on the classifier since the number of labels have been reduced significantly.

K-Nearest Neighbor is a discriminated classifier in which the prediction works by taking into consideration the observation around the location of the new data to be predicted. To clarify this concept, assume that we are planning to predict the class label of an input record using 5-Nearest Neighbor. In this example, once the new record is represented in the solution space, the 5-Nearest Neighbor works by determining the class label of the 5 nearest point to the new record. Then, a voting mechanism is applied and the new record is assigned to the label with the major points. The key performance factor of this classifier is the number of neighbors that are considered during the search (k). in this work, we have noticed that the performance of this classifier stabilizes once the k reaches to 15 (k = 15). Thus, in the experiment shows below, we have sat k to 15.

Table 1 illustrates the performance measures for this classifier. From the table, we can see that the K-Nearest Neighbor classifier performs slightly better than the Naïve Bayes model. This improvement (around 5%) in performance is related to the cleansing procedures that are applied in this work since it contributes towards eliminating unnecessary features and therefore reduce the dimensionality of the solution space. In addition, Table 1 shows the performance for only the header of the HS Code. From this table, it is clear that the accuracy of this classifier has been improved (71.72% accuracy) due to the above-mentioned factors similar to Naïve Bayes. Decision Tree classifier uses the input data to build a tree. This tree is eventually a conditional statement (if-else statement). In this mechanism, the node of the trees is selected based on the corresponding label. In other words, by starting from the root, the selection of the node will be based on the features and their impact on the class label. In each level, the features with the most impact on the class label will be selected as this particular level nodes. The behavior and performance of this classifier depends on the length of the tree and the number of labels. In our experiment it is not expected that this classifier to obtain significant classification behaviors due to the high number of labels. This is confirmed by the results which is shown in Table 1, where this classifier achieves 47.84% accuracy for the entire HS-code matching experiment. Random Forest is an ensemble learning technique in which the classification is performed based on the outcome of several learning models. In this work, the sub- learning model represents decision trees. By combining several decision trees into a random forest, the boundary of prediction is expected to be more accurate. And this is was clearly illustrated by the achieved results (66.21%) where the accuracy has improved significantly compared to the previously discussed Decision Tree classifier as shown in Table 1. Linear Support Vector Machine is an efficient well-known method that has been widely adopted to address several machine learning problems. Typically, a machine learning model tries to determine a hyperplane which divides the solution space according to the classes. For instance, assume that we have a binary problem, a hyperplane in this case is expected to divide the space such that one of the classes will be located to the right of the hyperplane and the other is located to the left of the hyperplane. In Linier Support Vector Machine, the optimization objective is to determine the location of this hyperplane such that the distance between the nearest point in each class and the hyperplane is maximized. This distance is typically referred to as a margin and therefore, the problem addressed in this situation can be denoted as maximizing the total margin. The results shown in Table 1 indicates that the linear Support Vector Machine achieves significant results than other classifiers in all of the cases (75.4%), where in the header only case, the improvement is much significant compared to the entire HS Code experiment (84.58%). This is related to the fact that in the header only case, the number of classes will be significantly lower and this reduces the complexity of locating an efficient hyperplane to separate the classes. In other situation where the experiment is performed to predict the entire HS Code, the complexity of the solution space increases the pressure on the classifier to obtain an efficient hyperplane locality.

Adaboost is an ensemble learning technique where the main idea of this classifier is to determine the weight of each sub-classifier based on its behavior in the overall evaluation. For instance, assume that Adaboost consists of three round optimizations. In each round, the weight of the classification factors depends on the outcome of the previous round of evaluations. Thus, this classifier aims to optimize the classification factors weights in iterative manner based on the behavior of the previous round. Unexpectedly, this classifier performance was relatively low compared to other machine learning classifiers as shown in Table 1. This is related to the fact that in our problem the number of features (after tokenization) is high. Therefore, Adaboost, failed to obtain an efficient classification behavior during the experiments which is bounded by 24 h running time.

4 Discussion

This study aims to investigate the problem of predicting the HS Code for the goods based on the users' input description. This prediction aims to reduce the revenue loss caused by using wrong HS Codes in order to reduce amount of duty that the trader is expected to pay to the customs administration. From the results, it is clear that the machine learning models can be used to help in predicting the HS Code for the user goods descriptions. The relatively low classification performance can be contributed by different factors such that the quality of the description and in some situations the planning of providing falsified descriptions. Since any machine learning model will be able to predict the description quality is bad due to poor English skills or due to the intention of the user to provide such low-quality description. The work that has been carried out to answer the research question which deals with the ability of using machine learning techniques to build HS Code Prediction Model establishes the ability to perform such tasks using machine learning techniques. To address the main objective question of this research, the dataset that is provided by Dubai Customs was used to predict the HS Code of the provided users' description using six machine learning models. The adopted machine learning models have achieved promising results specially the linier Support Vector Machine which achieved the highest accuracy of 76.3%.

5 Conclusion

This study underpins the challenges that confront Customs relating to the use of Harmonized System (HS) Code in their operations. Through the use of the case study approach and the development of a machine learning-based HS Code Prediction Model. Considering the benefits of using the HS code, there is subsequently a need to also address the issues it entails. The significance of this study is rooted on the interest of the research to contribute to empirical knowledge, especially in filling in gaps about the use of machine learning approaches in improving HS Code at Dubai Customs. As the results of the study reveal, the machine learning models are useful in predicting HS Code for the user goods descriptions. In addition, it is also noted that one of the six machine learning models used, the Linear Support Vector Machine, reveals a high accuracy of 76.3% in relevance with predicting HS Code using the dataset provided by the Dubai Customs.

However, this study is also not without limitations. Since a case study approach was adopted, the research is only limited to Dubai Customs as the machine learning models were only used on the dataset obtained from said government agency for predicting HS Code. On the other hand, information presented in the study can be used for future research, specifically on how gaps and other challenges in HS Code can be addressed among government agencies that use the modern technology. It can also be used to explore and identify strategies that will positively affect the efficiency of using HS Code in the public sector. The adoption of machine learning models as predictive frameworks and their respective effectiveness and/or efficiency can additionally be examined in future.

Due to the complexity of the problem, as a future step, we will explore the applicability of developing a divide-and-concur approach to address this problem. Whereas, hieratical prediction can be build starting from the HS Code header until identifying all of the HS Code subsections. Additionally, we are planning to explore the benefits of employing deep learning-based approaches to address this problem.

References

Alqaryouti, O., Khwileh, H., Farouk, T., Nabhan, A., Shaalan, K.: Graph-based keyword extraction. In: Shaalan, K., Hassanien, A.E., Tolba, F. (eds.) Intelligent Natural Language Processing: Trends and Applications. SCI, vol. 740, pp. 159–172. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67056-0_9

Alqaryouti, O., Siyam, N., Monem, A.A., Shaalan, K.: Aspect-based sentiment analysis using smart government review data. Appl. Comput. Inform. (2019, Accepted Manuscript)

Che, J., Xing, Y., Zhang, L.: A comprehensive solution for deep-learning based cargo inspection to discriminate goods in containers. IEEE Xplore (2018)

Ding, L., Fan, Z., Chen, D.: Auto-categorization of HS Code using background net approach. Procedia Comput. Sci. **60**, 1462–1471 (2015)

KPMG International: Transforming the tax function through technology (2018). https://home.kpmg/content/dam/kpmg/au/pdf/2018/transforming-the-taxfunction-au.pdf

Luppes, J.: Classifying short text for the harmonized system with convolutional neural networks. Master Thesis, Radboud University (2019)

Mohammed, M., Khan, M.B., Bashier, E.B.M.: Machine Learning-Algorithms and Applications. CRC Press, Boca Raton (2017)

United Nations: Globally Harmonized System of Classification and Labelling of Chemicals (GHS), 5th Revised edn. United Nations, Geneva (2013)

Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. **5**(4), 13–22 (2000)

Weerth, C.: Basic Principles of Customs Classifications under the Harmonized System. Global Trade Customs J. **3**(2), 61–67 (2008)

Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Elsevier, Burlington (2011)

Youyi, W.: Exploration of Data-Driven Intelligent Customs (2017). https://www.eiseverywhere.com/file_uploads/c7e054aa02ad13907d6ad513ea57b8d_session3-YouyiWu.pdf

Zang, B., Li, Y., Xie, W., Chen, Z., Tsai, C., Laing, C.: An ontological engineering approach for automating inspection and quarantine at airports. J. Comput. Syst. Sci. **74**, 196–210 (2008)