National College of Ireland

Title

MSc Research Project
Data Analytics

Edith Hernández
Student ID: X21198764

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rocio Edith Hernández Olguin
Student ID:	X21198764
Programme:	MSc. Data Analytics **Year:** 2023
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	December 14th
Project Title:	An Ensemble model to predict the classification of goods using text description.
Word Count:	6,881 **Page Count.** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 14 Dec 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	**x**
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	**x**
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	**x**

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Ensemble model to predict the classification of goods using text description

Edith Hernández
X21198764

Abstract

The main purpose of this research is to tackle the challenge of classifying 6-digit codes based on product descriptions. In order to achieve this, we will suggest an approach that combines NLP techniques, pre-trained word embedding and similarity search libraries.

There is a growing need, for effective methods to categorise products from large datasets, especially for customs authorities. The experiment intends to have the potential to improve the accuracy and efficiency of categorizing imported goods by leveraging advancements in Natural Language Processing (NLP) and deep learning. The research process will involve data collection, analysis, and experimental assessment. Every step is properly aligned with the CRISP-DM model.

Integrating FAISS in the proposed experiment improves the accuracy in using RoBERTa classification, which achieves 80%. The opposite case using FAISS and Distilbert classification got less than 1%.

The expected outcomes include gaining an understanding of the challenges and possibilities associated with classifying goods as well as developing a practical solution that can be applied in various contexts.

Keywords: *Natural Language Processing, HS code classification, RoBERTa, similarity search.*

1 Introduction

HS code classification, also known as the Harmonised Commodity Description and Coding System, is an internationally recognised numerical categorisation system that is employed for the purpose of classifying commodities in the context of global trade. Standardisation and consistency in product classification are crucial for customs administrations, enabling the effective identification and regulation of different products (Liya, et al., 2015). HS codes play a crucial role in facilitating customs administrations' determination of the suitable tariffs, taxes, and regulations applicable to both imported and exported goods. However, manual HS code classification can be a time-consuming, complex, and expensive process (He, et al., 2021). The

present research project ensures adherence to regulatory requirements and mitigates the risk of financial loss.

HS codes consist of a six-digit identification and are unique for each product. To classify all kinds of exported/products, the code is organised into 21 sections and 97 chapters with more than 5,000 product groupings in total and the World Customs Organisation (WCO), which is the entity responsible for managing the codes and updating them every five years[1].

For international classification purposes, the HS code is composed of at least six digits, of which the first two digits represent the Chapter, the next two digits identify the Heading, and the last two digits represent the subheading. The HS code may have more digits from seven to twelve, the length depends on the destination country for internal use of each country.

To illustrate how the HS code classification works, we take an example of the fruit apple, which is categorised under the HS code 080810, indicating the following:

- **08** represents chapter eight: Edible fruit and nuts; peel of citrus fruit or melons,
- **08** represents the heading eight: Apples, pears, and quinces, fresh,
- **10** represents the subheading: Fresh apples.

The code is internationally accepted, and every product must be classified correctly. Many small businesses struggle to classify their products correctly. To face the problem, they must invest money and time to get the correct code classification by paying a customs broker or hiring expensive services. The purpose of the research is to use NLP, which involves the use of computer algorithms to analyse, understand, and generate human language, interpret the "colloquial description" and relate with the "Official description" that HS code has and suggest the best approach to the product, as the volume of global trade increases, the use of NLP can help manage the growing amount of product classification data more efficiently and effectively (Zhang, et al., 2018). For HS code classification, the use of pre-trained word embeddings and similarity search to reach the main objective.

While conducting our research, we will encounter obstacles related to product description (Luppes, 2019) including but not limited to:

- Ambiguous description and the limited amount of information,
- The description does not follow a natural language structure,
- Misspelling and grammatical errors,
- Technical description including technical abbreviations,
- Multiple words, using synonyms, and abbreviations for the same products.

[1] http://www.wcoomd.org/- /media/wco/ public/global/pdf/topics/nomenclature/ activities-and-programmes/30-years-hs/ hs-compendium.pdf

The research question for this document: How can machine learning techniques in combination with NLP, improve the accuracy of HS code classification in the product classification?

To answer the research question, we use pre-trained embedding to NLP, so we can cope with some limitations that we can find in the product description once we have applied NLP techniques, by RoBERTa's pre-trained word embedding. We implement the RoBERTa model and similarly search to classify and get the most accurate 6-digit number based on the textual description.

The main contributions of this work are:

- Suggest a model by combining NLP, using RoBERTa pre-trained word embedding, and FAISS applied to HS code classification. With the research, we highlight its weaknesses and strengths in the research context,
- We share the datasets used in this research, which we got from the Department of Finance Bureau of Customs (Republic of the Philippines). The dataset consists of four files from November 2022 to February 2023 with 2,676,026 records.

This paperwork is structured as follows: Section 2 provides an understanding of the methodology and insights gained from previous applications related to the research topic. Section 3 explains the methodology used. Section 4 an overview of the HS code classification design process is presented, while in Section 5, we describe the proposed methods and models. Section 6 explains the process used and the results obtained. Finally, in Section 7, we conclude the experiments, share insights, and discuss future work.

2 Related Work

In recent years, HS code classification has been automated using machine learning approaches, derived from the importance that it has had in recent years in the emergence of borderless eCommerce and the facilitation of agreements between countries. In this state-of-the-art, we compare and critically evaluate several machine learning methods for classifying HS codes.

2.1 Conventional Machine learning techniques

A research study of machine learning for product classification based on textual descriptions was presented by (Lima, et al., 2020). The authors gave an overview of various machine learning methodologies used for product classification, including Naïve Bayes, Support Vector Machine (SVM), Decision Trees, and Deep Learning-based methods.

In related research, "A Commodity Classification Framework Based on Machine Learning for Analysis of Trade Declaration", the authors suggested a methodology for classifying commodities based on machine learning for the examination of trade declarations. The authors

proposed a method that relies on algorithms for text classification and logistic regression. The suggested strategy performs better than other established machine learning techniques like SVM and Decision Trees (He, et al., 2021).

To anticipate HS code, the authors conducted a comparative analysis of their suggested methodology with conventional machine learning models, including Naïve Bayes, K-Nearest Neighbour (KNN), Decision Tree, Random Forest, Linear SVM and Adaboost. SVM performed better in terms of accuracy than other machine learning algorithms, achieving a percentage of accuracy of 76.3%, according to the authors' experiment based on the Dubai Customs dataset (Altaheri & Shaalan, 2020).

An investigation using machine learning to categorise HS codes for apparel. In order to classify fashion products, the author compared Machine Learning methods such as using the Naive Bayes, the Multinomial Logistic Regression, the Decision Tree, and the Random Forest. The suggested method classified fashion items with an excellent accuracy of 57.66 (Barbosa, 2021).

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of neural network renowned for their exceptional ability to effectively process data organised in a grid-like structure, such as images or text. Taking advantage of this deep learning method, (Guo & Xiaodong, 2022) suggested a deep learning-based method. The classification of HS codes was done by the authors using a deep learning model based on CNNs. The empirical approach exceeded traditional machine learning techniques, achieving an accuracy of 87.5%.

Luppes (2019) described in their research, their approach to tackle short textual descriptions by proposing a CNN model for accurately classifying short text descriptions in the Harmonised System. The researchers found that the use of pre-trained word embeddings significantly improved the performance of the model, and he proposed a model architecture that incorporates pre-trained word embeddings and applies multiple filters of varying widths to capture different levels of n-gram features. The model was trained on a dataset of over 50,000 product descriptions and evaluated using various metrics achieving an accuracy of 92.4%. The suggested CNN model outperforms other well-known machine learning techniques, such as Naive Bayes and SVM, according to the author's research.

To capture both text and visual features of products, a suggested method blends CNNs with Long Short-Term Memory (LSTM) networks. Comparing it to existing methods, the proposed model achieves good classification accuracy on a sizable dataset of international eCommerce transactions. The paperwork outlines a deep learning-based solution to the issue of automatic HS code classification in cross-border eCommerce that makes use of both text and image characteristics. The authors emphasise the difficulties caused by the complexity of cross-border eCommerce transactions as well as the significance of precise HS code classification for customs clearance and supply chain management. Additionally, they examine current methods

for classifying HS codes and note a lack of studies on how to combine text and visual data to accomplish this goal. (Wu, et al., 2018)

A similar approach was seen in 2019, the authors suggested a hybrid method for classifying text that merged a CNN with a Recurrent Neural Network (RNN) to classify text data. In their model, local text features were captured using CNN, and long-term dependencies were captured using LSTM. The study was conducted in a comparative analysis by utilising two Chinese datasets, together with five English datasets, the authors tested their model and discovered that it performed better than other cutting-edge models. The method outperformed traditional machine learning techniques, developing an accuracy of 89.9%. (Wang, et al., 2019).

On one hand, a related study using a pre-trained language model suggests four-digit HS codes and retrieves key sentences from the HS manual that are most related to the product followed by suggesting six-digit HS codes using product descriptions and retrieved sentences. Their model was supported using KoELECTRA, which was specifically created to interpret and process words, equivalent to the cognitive processes involved in human reading and comprehension customised for Korean language tasks. The research presents the results of testing the model on recently examined electrical equipment (Chapter 85) and its successful application in the Korea Customs Service. The model outperforms the winning solution used in the product classification challenge in the eCommerce sector having an accuracy of 95.5% (Lee, et al., 2021). On the other hand, (Zhang, et al., 2018) introduced a more complex CNN architecture termed Text-CNN, with numerous convolutional and pooling layers, and produced cutting-edge outcomes on several datasets, their technique extracts information from product descriptions and forecasts the relevant HS code using a CNN and a LSTM network. The proposed approach outperformed conventional machine learning techniques, with an accuracy of 87.5%. Their suggested hybrid deep-learning classification model was limited to prediction in electronic customs clearance.

Provided a strategy for classifying import and export goods, the authors, combined a Hybrid CNN with an Auxiliary Network (HybridCNN-AN). The Shallow Structured CNN (SSCNN) component improves classification accuracy by addressing issues brought on by hierarchical categories and structured texts. The complexity of the model is characterised by employing three convolution operators for feature learning at various levels, which consist of their SSCNN, Deep Pyramid CNN (DPCNN), and Text CNN (TextCNN). Their model achieved superior classification accuracy compared to other models, having 92.33%. However, the study had some limitations. It may not apply to the customs systems of other nations because as focused on the Chinese system and the paper did not go into detail about how the model was implemented (Zhou, et al., 2022).

The classification of HS codes with transfer learning and pre-trained weights was the subject of another article. The research suggested the application of pre-trained weights in transfer learning boosts the categorisation of HS codes by focusing on the knowledge acquired from pre-existing deep learning models that have been trained on extensive datasets. The utilisation

of pre-trained weights allows the model to effectively use the acquired information from a distinct yet interconnected task or domain. This methodology enabled the model to effectively comprehend and capture the fundamental patterns and representations included in the textual descriptions of commodities. Different methodologies were examined, and it was found that the Google Universal Sentence Encoder exhibited strong performance when applied to English data, while the STSB-XLM-R-Multilingual model was effective for mixed foreign language data. Consequently, it facilitates the attainment of more precise and resilient categorisation outcomes for HS codes (Pain, 2021).

The incorporation of deep learning and digital image analysis in the HS code classification system enabled the identification of the appropriate HS code for coffee beans. The process involves the analysis of digital photographs of coffee beans, utilising deep learning algorithms to identify discernible patterns and characteristics that are indicative of particular HS codes. This functionality enables the system to identify the suitable tariff code for the coffee beans achieving an efficacy rate exceeding 80%, facilitating their import or export activities (German, et al., 2022).

In conclusion, Deep learning models have demonstrated favourable outcomes results when machine learning techniques such as CNN and LSTM have been applied to HS code prediction customs classification. Unfortunately, their limitation is the complexity of the models and the volume of training data required.

2.3 Textual Description Classification Using Machine Learning

In their investigation, the authors proposed a method for categorizing items using textual descriptions and sentence retrieval. Their method extracts pertinent data from product descriptions using sentence retrieval, which is subsequently applied to product classification. The approach outperformed conventional machine learning techniques, reaching an accuracy of 91.6% (Lee, et al., 2021).

Deep learning models like DNN and CNN showed that machine learning algorithms, especially, can correctly give HS-6 level codes based on written descriptions of goods. The study found that methods like weighted words, word embeddings, and lemmatisation made code assignments more accurate. This study demonstrated machine learning can be used to automate the process of assigning HS-6 codes. The analysis covered over one million US descriptions, demonstrating that it could make customs systems more accurate and efficient and support fraud detection, achieving 61% precision based on F-1 weights by the DNN model. (Ruder, 2020)

Sentence retrieval has been suggested as a strategy to extract pertinent information from product descriptions, and machine-learning techniques have been used to classify products based on textual descriptions. These works show how useful methods can be for text classification problems.

2.4 Word Embedding Methods for Text Classification

In their respective studies from 2022 and 2021, the authors employed the CNN algorithm for the purpose of text classification. In the field of text classification, CNNs have the ability to acquire knowledge regarding significant textual features, such as specific words or phrases. The primary purpose of its design is to effectively capture both the semantic and syntactic significance of words inside a given language. Word embeddings are commonly employed in the field of text categorisation to convert words into numerical vectors. By applying these obtained features, CNNs can effectively make predictions about the category or sentiment associated with the given text. The study revealed that the use of CNNs in conjunction with word embedding techniques such as Word2Vec, GloVe, and FastText as input or machine learning models can significantly enhance the precision of text classification. The Fast Text approach of word embedding achieves the highest categorisation accuracy. However, the precision the authors declared the effectivity of the model would depend on the datasets. (Merlin & Shini, 2021).

ALBERT, RoBERTa, and DistilBERT are relevant transformer-based models applied in many different NPL applications. The utilisation of an ensemble model enables the incorporation of both contextual information acquired by transformer models and local patterns detected by the convolutional layers of TextCNN. By combining these models, the ensemble model can use the knowledge that they each bring to make better classifications (Hua, et al., 2022).

The earlier studies covered in this section illustrate the promise of NLP and machine learning methods. There is still potential for improvement, notably in addressing the difficulties presented by confusing or irrelevant information in product descriptions (Liya, et al., 2015).

A potential disadvantage of both pieces of research is that they only examined a small number of word embedding techniques and did not investigate further approaches. On one hand, the scalability and generalizability of the models, the cost of computing, and the need for a lot of training data; on the other hand, a few approaches were limited to a specific chapter, such as electronic or fashion, or applied to a specific country. However, more complicated models often require more training data and computational resources. A limitation of the research was that none of the studies mentioned the computational resources used to execute models. Additionally, the integration of a human expert in the classification procedure was emphasised to guarantee reliability and mitigate any potential risks.

3 Research Methodology

In this study, we align the experimental study to the CRISP-DM model, which stands for the Cross-Industry Standard Process for Data Mining. This model offers a structured approach to data mining projects and is comprised of six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Figure 1.

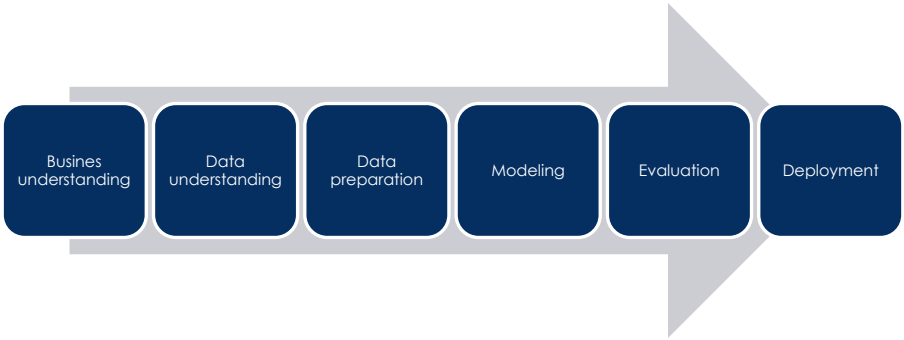

Figure 1: CRISP-DM model.

The first phase of the CRISP-DM model is called the business understanding phase, and it is responsible for defining the aims and objectives of the data mining project. The main objective of our project is to minimise the misclassification of goods and therefore the loss of income for the customs authorities and to reduce costs and have satisfied customers on the part of the trader by reducing delays and hidden costs associated with trade.

The second phase of the model called Data Understanding, involves the collection and analysis of data to obtain an understanding of the quality, integrity, and applicability of the data for the project objectives. The data will be analysed, identifying possible issues and findings.

Data Preparation is the next phase. It consists of choosing, cleaning, transforming, and preparing the data in order to facilitate the modelling process. Considering that the quality of the data that is used for modelling can have a big impact on the accuracy and effectiveness of the results, this phase is quite important.

During the fourth step, known as modelling, a number of different modelling techniques are applied to the data that has been prepared in order to develop a model to make predictions. The process of identifying acceptable modelling approaches, constructing, and evaluating the model to see whether it is effective and whether or not it is suitable for our business problem.

During the Evaluation phase, the model is evaluated based on new data, performance, and the results to be introduced into the decision-making process, the last phase of the model.

3.1 Business understanding

One of the important aspects of this study scope is to reduce the costs associated with incorrect classification of the items and increase the merchant confidence to sell their goods out of border by providing a good service to their clients. This misclassification is usually due to the merchant's ability to correctly determine the required HS Code.

3.2 Data Understanding

The datasets were gathered from the website of the Department of Finance Bureau of Customs (Republic of the Philippines), which was used for this research, the datasets were available on their website. The data consists of four files from November 2022 to February 2023, integrating 2,676,026 records. The main attributes are the HS Code and the textual description of the goods. This section will discuss the processing methods that we utilised to examine, analyse, and prepare the data for the model.

3.3 Data analysis

In the process of analysis, we have identified a few significant aspects that need to be considered during the process of developing the learning-based model. The quality of the good description and the number of expected classes (labels).

Regarding the good descriptions, for example, the spelling errors present in these descriptions, short descriptions using one or two words and similar words were used in more than one product. In relation to the number of classes, the data has 4,863 labels, the categories 84 and 82 most common in the dataset and 43 y 14 are the less common categories. Figure 2. Additionally, there have been observations made regarding significant variations in words linked to each class. For example, the HS Code 871410 has 29,177 records while 711011 has one record.

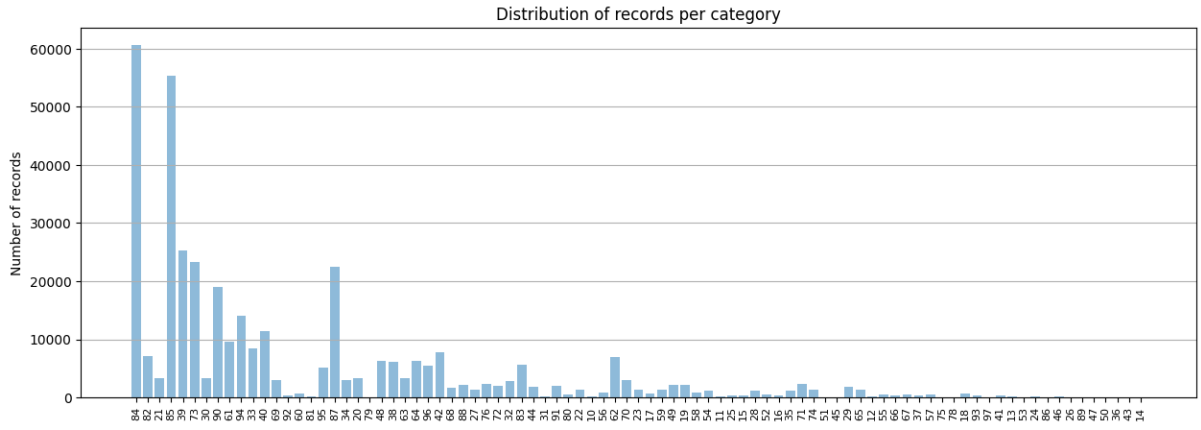

Figure 2 Overview of categories distribution.

Besides, the average number of words per textual description is 5.

3.4 Data Cleaning

In this section, we are going to describe the steps to execute in cleaning and pre-processing of the data. These steps can be summarised as follows:

- **Remove duplications**: The success of the machine learning model could be affected by having duplicate records, we found that 56% of the data were duplicated and we removed it from the dataset in the pre-processing step,

- **Remove punctuation and stop words:** Punctuation and stop words do not provide any significant value to the description of the goods, for that reason we have removed them,
- **Remove empty values:** As our main target is to find the best HS code based on the textual description, mistakes may be made because some of the values are null,
- **Remove numbers and measurement units:** For the purpose that we want to achieve, measurement units do not add useful information to our model. It is, therefore, we have to remove numbers and quantity measures such as ml, kgs pcs, etc.
- **Lemmatisation:** This process takes a word to its original form. Since words with the same root must be treated the same, using lemmatisation should lower the number of words we need to process and improve the performance.

Furthermore, in conjunction with the previous steps, the textual description has been transformed into lowercase letters, remove extra white spaces in order to have better data for the model. The output of this stage was carefully and randomly analysed in the interest of identifying the words that do not add any value to the item description.

3.5 Processing a sample

We chose a random sample of 500,000 as raw data for this work due to the limited processing power of the machine. After completing the cleaning processes and pre-processing, 362,494 were left as a result and 4,277 unique labels and memory usage of 16.6+ MB. These records will be used to train the RoBERTa and DistilBERT models.

3.6 Cleaning statistics

After executing cleaning and pre-processing methods, the following table presents statistical details after these processes.

Table 1. Statistical data details

Attribute	Count	Unique	Top	distribution
Hs code	362,494	4,277	871410	7,363
Description	362,494	351,161	part	31,294

On one hand, our categorical variables are “goodsdescription”, and “lemmadescription” where “goodsdescription” is the original textual description coming from the document and “lemmadescription” is the column created after pre-processing the information. This last column was the result of the pre-processing step and has been tokenised.

On the other hand, there are 3 numerical variables: “hscode”, “category”, and “totalwords”, where hscode is our target variable. Category presents the first 2 digits; this column will be used in step one of our models and total words contain the numbers of words that integrate our

textual description. To analyse the effects and investigate the reliability of the models, we employed basic descriptive information to furnish insights into the general performance.

Table 2. Descriptive information total words column

Descriptive info	total words
count	362,494
mean	5.762560
std	4.041804
min	1.000000
25%	3.000000
50%	5.000000
75%	7.000000
max	50.000000

With regard to our target variable "HS Code", in table number 2 we present the basic descriptive information after these processes.

Table 3. Descriptive information HSCODE column

Descriptive info	HSCODE
count	362,494
mean	693013
std	223479
min	100119
25%	481420
50%	840690
75%	853650
max	970510

Basic descriptive information of the Category column is shown in Table 4. We observe that category number 84 is the most common category in the dataset.

Table 4. Descriptive information of Category column

Descriptive info	
count	362,494
mean	69
std	22
min	10
25%	48
50%	84
75%	85
max	97

3.7 Splitting

The dataset will be separated into distinct subsets, namely the training and test sets. These subsets will be utilised for the purposes of model learning and evaluation. The proportion will be 70/30, being 70% for the training subset and 30% for model validation. 253,745 are part of the training subset and 108,749 are the testing subset.

3.8 Model selection

This experiment includes the implementation and evaluation of well-known designs: RoBERTa as category version and embedding layer, and similarly search for the usage of FAISS and Cosine Similarity blended with RoBERTa embedding. The preference of those models ends up primarily based on their verified efficacy in previous studies research. The forecast of the 6 digits of the product is based on the textual description as input. In order to achieve it, the experiment is divided into two parts: the first part is predicting the first two digits (Category) by the implementation of RoBERTa/DistilBERT classification, with assist of similarity search in conjunction with RoBERTa/DistilBERT embedding, and second part, merging the textual description with sentences retrieval of key phrases in the official tariff catalogue regarding the category, the version will expect the prediction of the rest four digits. Figure 3 illustrates the proposed solution.

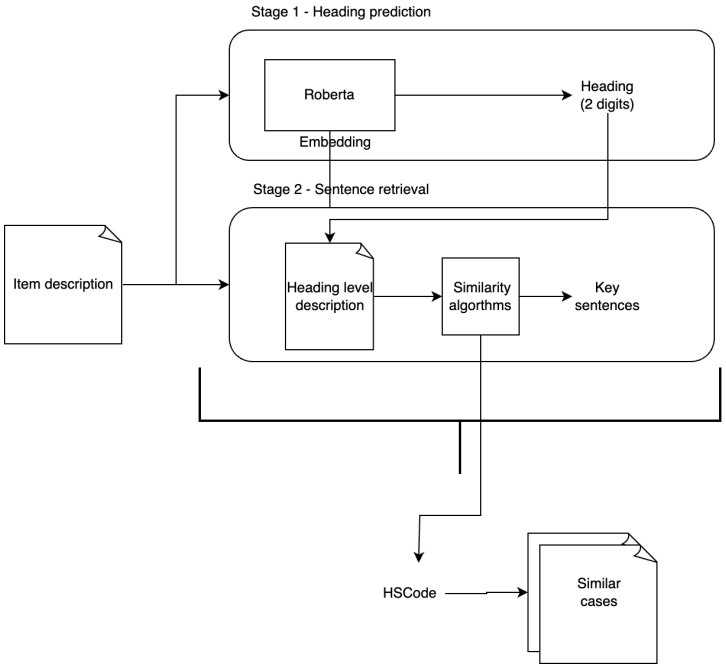

Figure 3 HS Code model prediction.

Experimental Setup

The experimental configuration consisted of deploying the models on the Colab research Google platform that changed into geared up with the requisite sources for conducting deep getting-to-know duties.

Category prediction.

As we mentioned before, the limited computational resources, we took a sample of 500,000 records. After the cleaning and pre-processing steps, we got as an input of our model 362,494 records where the number of unique labels is 4,277. Being the challenge as a multi-class classification problem, we used the RoBERTa and DistilBERT classification models from TensorFlow.

The model was trained in three epochs and eight batches, in each cycle the training loss decreased, and the accuracy increased, indicating an improvement in the ability of the model to fit the training data. The validation loss and accuracy provide insight into how well the model generalises to new and unseen data.

Table 5. Output training RoBERTa model

Epoch	Time	Loss	Accuracy	Val_loss	Val_accuracy
1	11989s 365ms/step	1.7268	0.5521	1.2568	0.6624
2	11936s 365ms/step	1.1859	0.6761	1.0977	0.7004
3	11938s 365ms/step	1.0151	0.7170	0.9995	0.7241

Table 6. Output training Distilbert model

Epoch	Time	Loss	Accuracy	Val_loss	Val_accuracy
1	7881s 241ms/step	1.6409	0.5710	1.1956	0.6777
2	7821s 239ms/step	1.0917	0.6494	1.0275	0.7155
3	7839s 240ms/step	0.9218	0.7402	0.9583	0.7355

3.9 Evaluation

For the implementation of this experiment, we used Python version 3.10, while the scikit learn, transformers and TensorFlow libraries were implemented for the model specifications. To evaluate the performance of each model we will assess the Confusion matrix, Precision, recall, F1-measure, and accuracy of the assessment metrics.

- **Precision**. It is a statistic for the classification model that indicates what percentage of all cases that were classified as positive were accurate interpretations. From a mathematical standpoint, it is shown as:

$$P = \frac{tp}{tp + fp} \qquad (1)$$

where *tp* corresponds to True Positive and *fp* stands for False Positives

- **Recall**. This metric indicates the percentage of all cases that were correctly identified as being absolutely positive. It can be expressed as follows:

$$R = \frac{tp}{tp + fn} \qquad (2)$$

where *tp* corresponds to True Positive and *fn* stands for False Negatives

- **F1-score.** It is a key metric used to evaluate the performance of a model. The quantity in question can be characterised as the harmonic mean of precision and recall, as expressed by the following equation:

$$F = 2 \times \frac{P \times R}{P + R} \qquad (3)$$

- **Accuracy.** The multi-class accuracy metric measures the proportion of properly predicted cases in relation to the total number of instances, resulting in representing the average accuracy of predictions (Jin & C.X., 2005). The formula expression for calculating multi-class accuracy:

$$accuracy = \frac{1}{N} + \sum_{k=1}^{|G|} \sum_{x:g(x)=k} I\big(g(x) = \hat{g}(x)\big) \qquad (4)$$

where *N* represents the number of observations, *G* includes all classes, and *I* is the function which returns a value of either one or zero.

4 Design Specification

To gain a better understanding of the proposed experiment, this document provides an overview of the implemented models.

RoBERTa (**R**obustly **o**ptimised **BERT a**pproach) created by Facebook is considered to be among the various self-training techniques that have achieved significant improvements in the field of language modelling, as expressed by the authors (Yinhan, et al., 2019). The study reveals that while BERT was notably undertrained, it demonstrates the capability to achieve performance levels that are on par with or surpass those of subsequent models that have been published.

DistilBERT is a reduced version of the BERT model. During the pre-training phase, the BERT model performed knowledge distillation, resulting in a 40% reduction in size. Despite this reduction, the model retained 97% of its language processing abilities and became 60% quicker.

The utilisation of RoBERTa for facts instruction represents a significant improvement for both small-scale organisations and large multinational organisations. This technique mainly focuses on integrating information for analysis functions as a way to extract valuable statistics[2].

The **FAISS** (**F**acebook **AI S**imilarity **S**earch) is a library that lets in similarity searches for multimedia documents. It surpasses the abilities of popular database engines like SQL by supplying nearest-neighbour seek implementations tailor-made for large-scale datasets, optimizing the trade-off between reminiscence utilisation, velocity, and accuracy. FAISS ambitions to deliver top-notch overall performance throughout numerous situations and consists of algorithms for looking in sets of vectors of any length. FAISS is usually accomplished in C, but it gives an optional Python interface and helps Graphics Processing Units (GPUs) through Compute Unified Device Architecture[3] (CUDA). CUDA platform and programming fashion, created by using NVIDIA, is designed for the purpose of famous computing on GPUs, emphasising parallel computing.

Cosine similarity is a metric in statistics retrieval and related research. This metric models a textual content record as a vector of phrases. By this version, the similarity between the two files can be derived by calculating the cosine price between the two files' time-period vectors. Implementation of this metric improves its capability to deal with semantics which means text by incorporating semantic checking between dimensions of two time-period vectors. This method aims to increase the similarity price between two term vectors which incorporate semantic relation between their dimensions with different syntax.

RoBERTa for Category Classification: RoBERTa is implemented as a classification model and embedding layer inside the first part of the test. This involves predicting the first digits of the HS code based on the textual description of the goods. RoBERTa, a highly proficient transformer-based model, demonstrates remarkable ability in identifying complicated contextual relationships within textual data.

In the second part of the experiment, we employed Similarity Search with FAISS and Cosine Similarity. The similarity search element makes use of FAISS, a high-performance similarity search library on one side to be compared with, Cosine Similarity performance in combination with the RoBERTa/DistilBERT embedding layer, leveraging its potential to generate meaningful vector representations of textual enter, which proves crucial for both category prediction and similarity search. These methods entail merging the textual description with sentence retrieval of key phrases from the official tariff catalogue related to the category anticipated by way of RoBERTa classification in the first part. This integration aids in predicting the remaining four digits of the HS code. This method refines the prediction to embody the closing four digits of the HS code.

[2] https://medium.com/analytics-vidhya/evolving-with-bert-introduction-to-roberta-5174ec0e7c82
[3] https://ai.meta.com/tools/faiss/

Algorithm Functionality. The algorithmic includes feeding the product's textual description into RoBERTa and DistilBERT for class prediction. Subsequently, the similarity-seek mechanism refines the prediction by considering relevant data from the official tariff catalogue. These embedding layers ensure that the version captures elaborate contextual functions important for accurate predictions.

5 Implementation

In this section, the recommended solution centred on a scientific approach oriented toward reaching forecasts of 6-digit product codes based totally on textual descriptions. The final section of the implementation focused on a couple of elements, which include RoBERTa/DistilBERT for categorisation, FAISS/Cosine Similarity for similarity search, along RoBERTa and BERT for embedding.

5.1 Model Integration

The RoBERTa model from the transformer's library changed into employed for the classification category. This involved predicting the first two digits of the HS code based on the textual description of the item. The key components from the transformer's library used include RobertaTokenizerFast, TFRobertaForSequenceClassification, DistilBertTokenizer and TFDistilBertForSequenceClassification.

The RoBERTa model turned into skilled over three epochs, each showing a decrease loss and an increase in accuracy, indicating improved model overall performance.
Outputs from this level encompass a skilled RoBERTa model capable of predicting the class of a product.

Similarity search combined with RoBERTa/DistilBERT embedding layer to further enhance the accuracy of predictions. The embedding layer generates relevant vector representations of the textual description input, important for each class prediction and similarity search.
By incorporating semantic family members amongst dimensions of term vectors, the model profits advanced capabilities in coping with semantics that means in the text.

FAISS, the performance similarity search library, turned out to be employed to leverage prediction by refining the search. FAISS was implemented with the use of the FAISS-CPU library and modified into used to carry out nearest-neighbour searches based on vector representations. This integration, allowed for the retrieval sentence of the official catalogue, contributing to the refinement of predictions. Finally, the aggregate of Similarity search and RoBERTa/DistilBERT embedding contributes to the accuracy of predicting the very last four digits of the HS code.

The accomplished experiments are developed at the Google Colab research platform with the important computational assets for the study.

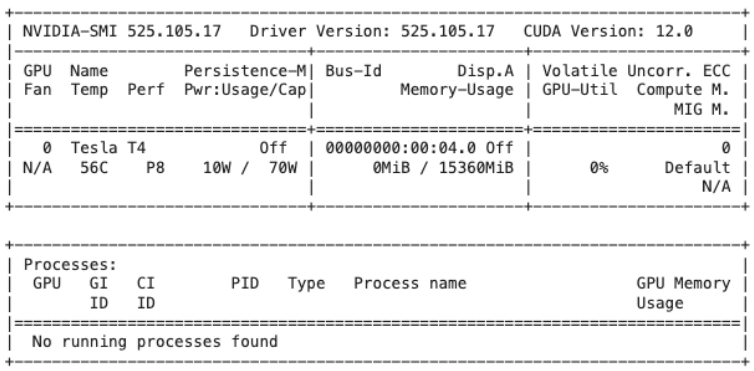

Figure 4. GPU Requirements used

5.2 Training and Evaluation

As mentioned before, trained the RoBERTa and DistilBERT classifications for category prediction over three epochs. Monitored key metrics which include education loss, accuracy, validation loss, and validation accuracy to metric model performance. In the evaluation metrics, we included confusion matrix, precision, recall, F1-score, and multi-magnificence accuracy in each model.

5.3 Tools and Libraries

Python version 3.10 served as the primary programming language for the implementation of the experiment. The main libraries implemented in the models are transformers, FAISS-CPU, and torch, which have been important for model specs and deep learning tasks.
The transformers library facilitated the mixing of RoBERTa and DistilBERT models for natural language processing functions. FAISS applied through FAISS-CPU, furnished similarity search skills tailored for huge-scale datasets.

6 Evaluation

Once the model is trained with the data, the RoBERTa model achieved an accuracy of 72%, meaning that it predicted the category correctly in comparison with DistilBERT, which achieved 73%. Predicting the HS Code to 6 digits RoBERTa got 80% in comparison with Distilbert 0.24%. The F1 score metric is also reported as 80%, which means it strikes a good balance between accuracy and recall, in our case, it was reported as 80%, meaning that the model correctly identified 80% of the truly positive cases. Finally, the precision metric was reported to be 80%, meaning that if the model predicts a positive square, it is correct 80% of the time. In summary, the RoBERTa model obtained reasonably good performance in the multi-class classification task, with balance and accuracy with recall, and an overall accuracy of 80%.

Table 7. Model evaluation results (First 2 digits)

Model	Accuracy	F1 Score	Recall	Precision	Time	Records
Roberta-base	72.41%	72.41%	72%	72%	10 hours	362,494

distilbert-base-uncased	73.55%	73.55%	73.55%	73.55%	5 hours	362,494

Table 8. Model evaluation results (6 digits)

Model	Accuracy	F1 Score	Recall	Precision	Time	Records
Faiss + Roberta	80.26%	80.26%	80.26%	80.26%	1 hour	10,875
Faiss + distilbert	0.24%	0.24%	0.24%	0.24%	1 hour	10,875

In terms of predicting the 6-digit HS Code classification, RoBERTa obtained 80%, which means that the classification model in combination with a similarly search significantly improved the accuracy.

6.1 Discussion

Using RoBERTa for category prediction offered promising results. The model, which was able throughout three epochs, continually showed improvement with the aid of lowering loss and incrementing accuracy. This version confirmed the ability to predict the preliminary two of six digits of the HS code, achieving 74% accuracy DistilBERT 73.55% in predicting the Category.

Regarding predicting the 6 digits of HS Code classification RoBERTa got 80% versus less than 1% of Distilbert classification.

Integrating FAISS for similarity search stepped forward category predictions. The version's capacity to carry out nearest-neighbour searches, facilitated via FAISS, extensively improved prediction accuracy.

Similarity search with the RoBERTa/BERT embedding layer in addition to great-tuned predictions, particularly in predicting the remaining 4 digits of the HS code. The embedding layer's capacity to generate significant vector representations, coupled with the semantic awareness of Similarity search. The combination improved the model's capability to comprehend the semantic relations within textual statistics.

The findings align with the effectiveness of transformer-primarily based fashions like RoBERTa in NLP tasks, as observed in preceding studies (Yinhan, et al., 2019). FAISS's position in optimizing nearest-neighbour searches for big-scale datasets resonates with its contribution to literature. The blended use of Cosine Similarity and embedding layers mirrors the success reported in leveraging contextual records in text information (Jacob, et al., 2019).

The performance of the model was optimal in general; however, human intervention is still necessary to corroborate that the HS code has been accurately predicted.

7 Conclusion and Future Work

Incorporating additional models by experimenting with opportunity transformer-primarily based fashions and similarity search techniques could provide insights into their comparative efficacy.

Even as the experiments showcased great fulfilment, in particular in the collaboration of RoBERTa, FAISS, and Cosine Similarity, there are opportunities for refinement. Acknowledging barriers and constructing upon recognised strengths lays the muse for destiny improvements in predicting HS codes primarily based on textual descriptions.

By considering an analysis of the errors, we will be able to have a better knowledge of the model improvement and opportunity areas that we may concentrate on in future research.

It is quite clear that this task requires more computational resources. In this experiment we tried to use Amazon Web Services, a p2.xlarge instance with 1 GPU and 4 vGPU was not enough to optimally execute the classification task. By executing 1 epoch, it would take 88 hours. As this kind of classification is a complex task, it is necessary to get more technological resources. We propose to experiment using a p3 instance, it could develop faster training for the classifiers.

References

Altaheri, F. & Shaalan, K., 2020. Exploring Machine Learning Models to Predict Harmonized System Code. pp. 291-303.

Barbosa, I. M. A., 2021. Using Machine Learning to classify HS codes for Fashion Products.
Chunrong, G. & Xiaodong, Z., 2022. *Research on Intelligent Customization of Cross-Border E-Commerce Based on Deep Learning,* s.l.: Mathematical Problems in Engineering.

German, C.-S., Irving, H.-V., Elías, R. & Karina, G.-F., 2022. *Automatic Tariff Classification System using Deep Learning,* Mexico: International Journal of Advanced Computer Science and Applications.

Guo, C. & Xiaodong, Z., 2022. Research on Intelligent Customization of Cross-Border E-Commerce Based on Deep Learning.. *Mathematical Problems in Engineering 2022.*

He, M. et al., 2021. A Commodity Classification Framework Based on Machine Learning for Analysis of Trade Declaration. *Symmetry. 2021,* p. 13(6):964.

Hua, C. et al., 2022. TextCNN-based ensemble learning model for Japanese Text Multi-classification. *Computers and Electrical Engineering,* pp. 1-12.

Jacob, D., Ming-Wei, C., Kenton, L. & Kristina, T., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,* Minneapolis, Minnesota: Proceedings of NAACL-HLT 2019.

Jin, H. & C.X., L., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering,* pp. 299-310.

Lee, E. et al., 2021. Classification of Goods Using Text Descriptions With Sentences Retrieval.

Lima, T., Nunes, F. & Oliveira, E., 2020. *Machine Learning for Product Classification based on Textual Descriptions: A Literature Review,* s.l.: In Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021).

Liya, D., ZhenZhen, F. & DongLiang, C., 2015. Auto-categorization of HS code using background net approach.. *Procedia Computer Science, Volume 60,* p. 1462–1471.

Luppes, J., 2019. *Classifying Short Text for the Harmonized System with Convolutional Neural Networks (thesis),* Kenya: s.n.

Merlin, S. D. & Shini, R., 2021. Comparison of word embeddings in text classification based on RNN and CNN. *Materials Science and Engineering,* p. India.

Pain, K., 2021. *HARMONIZED SYSTEM CODE CLASSIFICATION USING TRANSFER LEARNING WITH PRE-TRAINED WEIGHTS,* Nova Scotia: s.n.

Ruder, D., 2020. *Application of Machine Learning for Automated HS-6 Code Assignment,* Tallinn: TALLINN UNIVERSITY OF TECHNOLOGY.

Wang, R. et al., 2019. Convolutional recurrent neural networks for text classification.. *In 2019 International Joint Conference on Neural Networks (IJCNN),* pp. 1-6.

Wu, J., Peng, J. & Yu, J., 2018. Automatic classification of HS codes in cross-border e-commerce using deep learning. *2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI),* pp. 216-221.

Xi, C., Stefano, B. & Marko, v. E., 2021. Neural Machine Translation for Harmonized System Codes prediction. *ACM DL.*

Yinhan, L. et al., 2019. *A Robustly Optimized BERT Pretraining Approach,* Seattle, WA: University of Washington.

Zhang, W., Liu, Y. & Zheng, Y., 2018. Research on deep learning-based classification model of HS code in electronic customs clearance. *Energy Education Science and Technology Part A: Energy Science and Research, 36(6),* pp. 6356-6363.

Zhou, C. et al., 2022. Harmonized system code prediction of import and export commodities based on Hybrid Convolutional Neural Network with Auxiliary Network. *Knowledge-Based Systems.*