ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Similarity versus Supervision: Best Approaches for HS Code Prediction

Sédrick Stassin[1*], Otmane Amel[1*], Sidi Ahmed Mahmoudi[1] and Xavier Siebert[1 †]

University of Mons - ILIA Unit
Mons - Belgium

Abstract. With growing e-commerce flows and new legislative rules, customs representatives confront serious liabilities when completing customs declarations for their clients. In the latter, the Harmonized System (HS) code is a crucial component using 10 digits (HS10) to classify products and define national tax rates. In this paper, we first compare the performance of sentence embedding models using semantic similarity, and second, we assess the effectiveness of supervised models, both aimed at predicting up to the HS10 code. To the best of our knowledge, there is currently little research being conducted on this topic. We demonstrate the differences and respective strengths of each approach. Our results show the outstanding performance of the semantic similarity approach with a top-3 and top-5 accuracy of 89% and 94.8% respectively for HS10 prediction.

1 Introduction

Changes in legislation, increase in e-commerce flows, and customs controls are all contributing factors to the increased risk faced by customs representatives. In addition, one of the crucial elements in customs declaration that exacerbate this risk is the use of Harmonized System (HS) codes which are often caused by human mistakes. The HS code is composed of 6 international digits (HS6) and 4 other digits (HS10) that help to assign, at the European and national level, the tax rate of each product. The 10-digit structure is broken into groups of two digits, each of which is attributed to a textual description that provides a more detailed definition of the product group[1]. Today, artificial intelligence (AI) models can help customs representatives mitigate the risk encountered, by inspecting the classification of products (HS codes) in their declarations. This study compares two approaches to predict HS codes at level 6 (HS6), 8 (HS8), and even 10 (HS10): pre-trained sentence embedding models using semantic similarity, contrasted with supervised neural networks using the same sentence embeddings. We analyze the real-world feasibility of both approaches via performance and advantages. In the upcoming sections, we introduce related work in the field of HS code prediction, followed by our methodology for both approaches and our experimental results. Finally, we draw conclusions and offer suggestions for further research.

*These authors contributed equally to this work.

†The authors thank the support of the Infortech institute and the E-origin project funded by the Walloon Region within the pole of logistics in Wallonia.

[1]Belgian governmental free access database of HS code nomenclature: https://eservices.minfin.fgov.be/extTariffBrowser/browseNomen.xhtml?suffix=80&lang=EN

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2 Related Works

In this Section, we present studies related to the field of HS code prediction. He et al. [1] attempted to generate a four-digit prediction using data (name and description of the goods) associated with chapter 84. On two Chinese datasets of 220,000 and 8,900 samples, they achieved 88 and 99% accuracy using Bert [2] followed by either classification layer or CNN, or a symmetrical fusion between the two preceding models. The work of Luppes [3] gathered 79 million sentences from private and public sources to retrain the embeddings of the Word2Vec model [4]. The training of a one-dimensional CNN for the final HS code prediction problem using datasets of 13 million declarations resulted in 92% of F1-score for HS2, and 88% accuracy for HS4. However, the dataset labels used were not validated by experts, which does not guarantee the model's efficiency during inference. With features associated to fashion products (fabric, material), Barbosa et al. [5] developed a hierarchical model for HS code prediction. They created three distinct classifiers to predict the chapter (HS2), the heading (HS4), and the subheading (HS6) of the HS codes preceded by a rule-based one-hot encoding of the features. The first approach used the classifier to forecast the next HS code ramification utilizing only real data input, and it outperformed the second approach, which attempted to predict further ramifications of HS codes based on the initial prediction (HS2). Employing unbalanced data solutions (undersampling, oversampling) on machine learning models, the decision tree approach achieves the greatest performance with 88%, 67%, and 58% accuracy for two, four, and six-digit predictions, respectively. Unlike supervised techniques, Pain [6] used pre-trained Semantic Textual Similarity (STS) models based on sentence-level embedding to obtain an unsupervised recommendation of top-k HS codes (at any level). The recommendation is based on the cosine similarity of a provided data description to existing HS code descriptions in legal databases. From four datasets (ranging in size from 4,000 to 12,000) obtained through Electronic Data Interchange (EDI) files, the results showed a top-10 precision that varied from 43% to 70% on the best sentence embedding model Universal Sentence Encoder (USE), along with variations of the Bert model [2]. Closer to our work, Spichakova and Haav [7] introduced an innovative way that allows for the prediction of HS6 codes using a combination of two different measures on the Bill of Lading Summary 2017 dataset. First of all, using the Doc2Vec [8] embedding a cosine similarity measure of sentences describing a product is calculated, displaying the most similar text sentences associated with an input and the corresponding HS6 codes. In addition, a semantic similarity measure compared the actual HS code and the predicted HS code values based on the taxonomy of the HS code. The combined measure lets the predicted HS code be rated. our work introduces several contributions as follows: 1) we delve into the prediction of HS10 codes, an area where, to the best of our knowledge, limited research has been undertaken. 2) We provide a comparative analysis between the remarkable performance of similarity-based models and supervised models using pre-trained sentence embedding models and showcase their advantages.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

3 Methodology

In this Section, we describe our approaches for HS code recommendation. The first approach based on semantic similarity aims to find the appropriate sentence embedding model able to capture semantic relationships between commodity descriptions, along with the adapted distance metric (see Fig. 1). To further formulate this problem, let N be the number of declarations already validated by customs operators, and a mapping function $f(d) = Z_{input}$ that takes as input a new invoice description d and gives the embedding vector $Z_{input} \in \mathbb{R}^m$ where m denotes the dimension in an embedding space. We apply the same function over the N descriptions and we use a similarity metric g to obtain semantic distances between descriptions to draw a list of the k closest descriptions semantically. The second approach is a basic supervised neural network that takes as input Z_{input} produced by the same embedding models used for the first approach followed by a fully connected classification layer to obtain predictions on a given HS code level (e.g. HS6, HS8, HS10). The objective is to compare the strength of both approaches to draw meaningful conclusions.

4 Experiments

4.1 Dataset

The dataset comprises 95,903 customs declarations from e-Origin[2], encompassing 967 distinct HS6 codes, 1,181 HS8 codes, and 1,196 HS10 codes. The data was preprocessed by removing punctuation, special characters, and digits from text columns. Miswritten or combined words were addressed using a combination of WordSegment tools and a Python-based spell-checker.

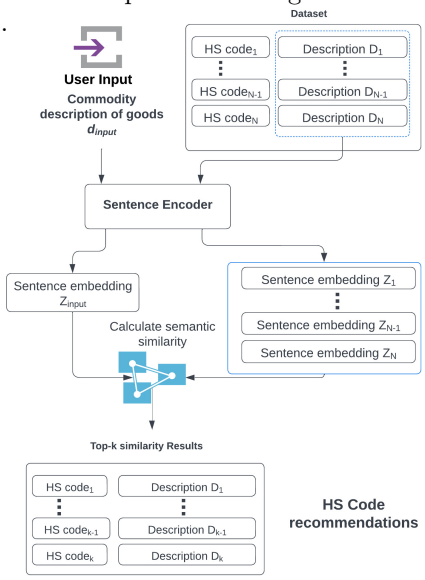

Fig. 1: Semantic similarity approach for HS code recommendation.

4.2 Setup

We evaluated multiple pre-trained sentence embedding models based on the Bert transformer architecture [2]: MPNet [9], MiniLM [10], USE [11], Roberta [12], DistilBERT[13]. Semantic similarity was computed using standard distance metrics: Euclidean, Cosine, and Manhattan distance, in addition to Chebychev and Minkovski for the sake of comparison. The models were implemented using Pytorch and executed on GPU resources. In the supervised approach, 40 epochs were conducted, partitioning the dataset into 80% for training, 10% for testing, and 10% for validation. A batch size

[2]https://eorigin.eu/

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

of 16 and a learning rate of $1e^{-3}$ were employed. Results were reported for models trained with varying minimum samples per class: 50, 400 to 600. This investigation sheds lights on the influence of the number of classes on supervised model performance. The classification layer includes a 20% dropout mechanism followed by ReLU activation, and the number of classes is determined by the previously set minimum samples (refer to Table 2).

4.3 Discussion

In the field of HS code prediction, it is crucial to acknowledge that a concise text description might correspond to multiple HS codes. Thus, empowering users to make informed decisions through a top-3 or top-5 classification, rather than an automatic top-1 choice, becomes more interesting. An automatic top-1 classification yields the most common option, which often falls short of yielding optimal outcomes. As demonstrated in Table 1, the semantic similarity-based approach exhibits remarkable HS code recommendation performance across all levels, achieving top-3 accuracies of 90%, 88%, and 88% for HS6, HS8, and HS10, respectively. Notably, this is realized without supervised training. Among the models assessed, MiniLM [10] consistently outperforms others, particularly in top-3 and top-5 accuracy when employing specific weights [4] for HS6 and HS8. Similarly, MPNet [9][8] excels in top-3 and top-5 performance for HS10. While differences in accuracy among distance metrics for a given model are slight, overall, Euclidean and Manhattan distances exhibit superior performance regardless of the number of digits, offering flexibility for recommendation tasks. Several noteworthy points arise from the analysis. Firstly, a substantial performance leap occurs between top-1 (around 62%) and other top-k accuracies, indicating the significance of results beyond top-1 and emphasizing the advantageous nature of semantic similarity. Secondly, the model showcases resilience to HS code complexity, with only a minor decrease when transitioning from HS6 to HS10 (from -2.5% for top-1 to -1% for top-5). Thirdly, this method facilitates recommendations for every existing HS code, a crucial edge over the supervised approach restricted by predefined class counts during training. Nonetheless, models reliant on semantic similarity do face limitations tied to historical databases. They may struggle to predict HS codes for novel products absent from historical data, curbing their efficacy. Yet, this challenge is not exclusive to semantic similarity models, and updating historical data is easier than retraining a supervised model. Shifting the focus to supervised learning, the outcomes are presented in Table 2. Across all HS levels, a marked performance drop is observed when training models with fewer minimum samples per class (e.g., from 600 to 50). While achieving nearly 30% top-1 accuracy with 209 classes categorized, the similarity-based approach attains over 60% top-1 accuracy with 1,196 classes for HS10, showcasing its superiority and utility over the supervised method. This notion is confirmed by previous research [1, 3], which confines supervised models to predicting limited HS codes or levels. Finally, it is worth noting that supervised models fail to detect instances outside known classes. In such cases, one can rely on high prediction values or develop a new binary supervised

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

model to distinguish between known and unknown classes. However, this complexity diminishes the practicality of the supervised approach compared to the similarity-based approach in real-world scenarios.

# of digits	Top-k			Model	Metric
	k=1	k=3	k=5		
6	**0,647**	0,900	0,956	MiniLM[3]	Chebyshev
	0,644	**0,910**	**0,961**	MiniLM[4]	Euclidean
	0,644	0,908	**0,961**	MiniLM[4]	Minkowski
8	**0,621**	0,878	0,950	DistilUSE[5]	Manhattan
	0,620	**0,889**	**0,952**	MiniLM[4]	Euclidean
	0,620	0,887	0,951	MiniLM[4]	Minkowski
10	**0,621**	0,883	0,947	DistilRoberta[6]	Manhattan
	0,618	0,880	0,947	MiniLM[7]	Manhattan
	0,617	**0,890**	**0,948**	MPNet[8]	Cosine

Table 1: Top-1, top-3, and top-5 accuracy of the semantic similarity-based approach with respect to the sentence embedding model and distance metric used.

# of digits	Min. data / class	# of classes	Top-k			Model
			k=1	k=3	k=5	
6	600	13	0.838	0.930	0.933	MiniLM[3]
	400	34	0.575	0.649	0.718	
	50	209	0.302	0.319	0.330	
8	600	13	0.844	0.885	0.918	MiniLM[9]
	400	33	0.633	0.662	0.664	
	50	209	0.282	0.294	0.316	
10	600	13	0.895	0.926	0.928	MiniLM[3]
	400	32	0.629	0.703	0.710	
	50	209	0.336	0.353	0.364	

Table 2: Top-1, top-3, and top-5 accuracy of supervised models with respect to the minimum number of samples per class.

5 Conclusion

Customs classification, particularly HS code assignment, poses a significant challenge for customs representatives. This study compares supervised neural net-

[3]weights: paraphrase-multilingual-MiniLM-L12-v2
[4]weights: multi-qa-MiniLM-L6-cos-v1
[5]weights: distiluse-base-multilingual-cased-v2
[6]weights: sentence-transformers_all-distilroberta-v1
[7]weights: all-MiniLM-L12-v2
[8]weights: paraphrase-multilingual-mpnet-base-v2
[9]weights: sentence-transformers/all-MiniLM-L6-v2

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

works and history-based models using semantic similarity. Both models employ identical sentence embedding models for HS code prediction up to 10 digits. Experimental findings underscore the similarity-based method's prowess using the MPNet [9] sentence embedding model with cosine similarity, achieving a remarkable top-3 accuracy of 89% and top-5 accuracy of 94.8% across 1,196 classes for HS10 prediction. For HS6 and HS8 predictions, the MiniLM [10] model excels. Results demonstrate the superior performance of similarity-based models over supervised models (35.3% HS10 top-3 accuracy with 209 classes). Moreover, the addition of new data and classes to the historical dataset enables the prediction of these classes, a capability lacking in the supervised model. As the number of classes increases, the performance of the supervised model diminishes significantly, revealing its limitations in real-world scenarios when faced with out-of-distribution classes. Future works could explore the application of semantic similarity to multimodal customs datasets.

References

[1] Mingshu He, Xiaojuan Wang, Chundong Zou, Bingying Dai, and Lei Jin. A commodity classification framework based on machine learning for analysis of trade declaration. *Symmetry*, 13(6):964, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Jeffrey Luppes, Arjen P de Vries, and Faegheh Hasibi. Classifying short text for the harmonized system with convolutional neural networks. *Radboud University*, 2019.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Ivânia Maria Afonso Barbosa. Using machine learning to classify hs codes for fashion products. 2021.

[6] Koustav Pain. *Harmonized System Code Classification Using Transfer Learning with Pre-Trained Weights*. PhD thesis, 2021.

[7] Margarita Spichakova and Hele-Mai Haav. Using machine learning for automated assessment of misclassification of goods for fraud detection. In *International Baltic Conference on Databases and Information Systems*, pages 144–158. Springer, 2020.

[8] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[9] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[10] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.