Text2Trade: A Semantic Search System with Monte Carlo Dropout Uncertainty Quantification for HS Code Retrieval

Meenu Ravi
The MITRE Corporation
McLean, Virginia, USA

Dr. Jacob Howard
The MITRE Corporation
McLean, Virginia, USA

Grace Kenneally
The MITRE Corporation
McLean, Virginia, USA

Ari Fleischer
The MITRE Corporation
Bedford, Massachusetts, USA

Abstract

The efficient and accurate identification of Harmonized System (HS) codes is crucial in the import and export process of commodities for international trade. This six-digit globally standardized numerical classification is important for describing products, assessing tariffs, and conducting trade analysis. However, assigning HS codes with speed and accuracy is challenging due to the complex hierarchical structure, large and dynamic set of codes, and subtle distinctions in product terminology. Recent advances in machine learning and information retrieval have improved HS code retrieval; however, existing approaches often struggle to capture semantic meaning about products, process long product descriptions, and remain largely proprietary in nature. To address these challenges, we present Text2Trade, a semantic search system that recommends relevant HS codes based on product descriptions. Our approach fine-tunes a bi-encoder with Multiple Negatives Ranking Loss (MNRL) and integrates Monte Carlo dropout (MCD) for both reranking and uncertainty quantification, trained on diverse and publicly available trade data. Through comprehensive experiments and an ablation study, we demonstrate that Text2Trade consistently outperforms several baseline models, achieving a Recall@10 of 97.8%. These findings highlight the system's potential to enhance the efficiency of trade policy analysis, helping the broader trade and policy community.

CCS Concepts

• **Information systems** → **Top-k retrieval in databases**.

Keywords

Semantic search, Retrieval ranking model, Neural information retrieval, Monte Carlo dropout, Uncertainty quantification, Sentence-BERT, Tariff classification, International trade

CHIIR '26, Woodstock, NY

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The World Customs Organization (WCO) has implemented the HS code system as a uniform way to categorize internationally traded commodities using standardized 6-digit codes and descriptions.[1] Each description provides detailed information that allows importers and exporters to accurately identify goods. The code-description pairs are organized hierarchically into groups of two digits (i.e., chapter, heading, and subheading respectively), as illustrated in Figure 1. Accurate HS code assignment determines import tariffs and regulatory measures, ensures compliance with trade rules, supports trade analysis, and prevents costly fines.

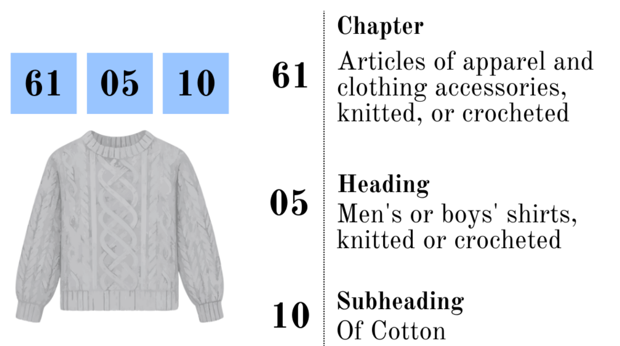

Figure 1: Example of a 6-digit HS code assigned to a product.

Despite its importance, HS code assignment remains largely a manual process carried out by domain experts, making it expensive and error-prone [5, 16, 31]. Reports have indicated that nearly 20% of shipments entered the country with incorrect tariff codes, leading to an estimated $21 million in annual revenue losses for the government. Furthermore, importers claimed back $136 million in refunds after discovering favorable misclassifications [16]. These figures highlight the financial consequences of HS code misassignment for both governments and traders.

This creates a challenging information retrieval (IR) problem as users must effectively search through thousands of hierarchically-structured codes and distinguish semantic nuances in product descriptions covering more than twenty sectors. For example, shirts alone maps to multiple HS codes under heading 6105, distinguishing man-made fabrics, textile materials, and derivative products. This complexity in terminology creates several challenges: commonly used products must be matched to formal classifications

[1]While countries often use extended tariff codes of 8–10 digits, the 6-digit level is internationally standardized and consistent across countries.

(e.g., "automatic data processing machines" as computers) while products with dual functionalities (e.g., smartwatches as both timepiece jewelry and computing devices) require nuanced interpretation to determine the appropriate code. For such ambiguous and complex classifications, effective search systems that narrow the relevant options with uncertainty measures can support expert human decision-making, allowing users to evaluate a small subset of relevant candidates rather than manually traversing the entire hierarchy.

While prior research in HS code retrieval has applied machine learning (ML) and deep learning (DL) methods [1, 5, 7, 8, 14], existing methods struggle with capturing semantic information, adapting to WCO's five-year revisions, enumerating the full range of options. Additionally, existing search tools are largely proprietary, limiting accessibility and often rely heavily on the user to correctly determine granular product characteristics, which may be unknown and time-consuming. While semantic search has been promising in general domains, there is limited research on addressing nuances in domain-specific terminology, and effectively representing model uncertainty to users in trade regulation domains where misclassifications carry significant financial and policy ramifications.

To address these challenges, we present Text2Trade, a semantic search model designed to improve the effectiveness and speed of HS code search. The main contributions of the paper are summarized as follows:

- We develop a semantic search engine that accelerates HS code search by fine-tuning a bi-encoder model with MNRL, on diverse and publicly available trade data.
- We conduct a comprehensive experimental evaluation of Text2Trade against six baseline models using real-world commodities, demonstrating consistently higher accuracy and significantly faster inference times.
- We integrate MCD to not only improve ranking but also provide interpretable uncertainty measures that help domain experts navigate ambiguous product descriptions.

2 Related Works

2.1 HS Code search tools

Recent works have tried to address manual HS code retrieval. The United State Census Bureau (USCB) provides a commodity search tool [3] as a more intuitive way to classify products through an interactive interface. However, such tools heavily rely on users to know exact product attributes such as composition, functionality, and size, some of which can be ambiguous or unavailable in real-world trade documentation. Not only do these tools struggle with processing long product descriptions, but their manual input process restricts users to submitting one query at a time, making them impractical for high-throughput applications across large product catalogs. Additionally, prior work in trade compliance highlight traditional rule-based and keyword search systems [25]. While these tools perform well on product descriptions that are direct and short, they struggle with vague or domain-specific terminology.

2.2 Neural retrieval approaches for HS codes

There has been significant interest in using ML and DL approaches to classify HS codes based on a product descriptions [5, 7, 14]. Specifically, Chen et al. [5] reformulated the HS code assignment task into a classification task by using an attention-based neural machine translation to capture the hierarchical structure of HS codes. Liao et al. [14] proposes a classification model using Bi-directional Long-Short-Term Memory Network (BiLSTM) to handle sequential dependencies on trade terminology with Enhanced Representation through Knowledge Integration (ERNIE) pre-trained embeddings and multiscale attention mechanisms. However, these approaches struggled with capturing specialized trade terminology.

2.3 Uncertainty-aware retrieval methods

While uncertainty quantification is widely used in natural language processing (NLP) problems [15, 30], it is underexplored in HS code search tasks. Though recent works in uncertainty quantification for NLP tasks, which include Monte Carlo dropout [2, 10, 13, 18, 24, 26] and Dirichlet-based Uncertainty models (DBU) [17, 21, 32], are prevalent, these techniques have not been systematically applied to domain-specific search tasks or been used as a method of reranking. We adopt MCD due to its efficiency and adaptability with bi-encoder models.

Therefore, to further improve the search of HS codes, our semantic search system, Text2Trade, is fine tuned on publicly available and diverse trade-related data spanning decades, taking into account the hierarchy. It achieves improved ranking quality and domain generalizability, while maintaining efficient inference speed. We further integrate MCD as both a reranking mechanism and uncertainty quantification method, allowing users to better handle ambiguous queries, while preserving human judgment in final code assignment decisions.

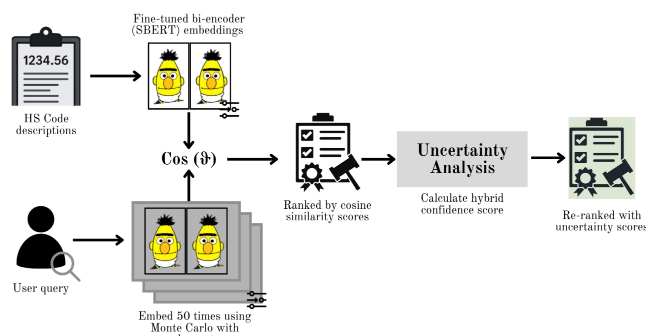

Figure 2: Design of Text2Trade semantic search with uncertainty quantification framework

3 Methodology

The overall framework of Text2Trade is shown in Figure 2. Given a text description of a product, the system retrieves semantically similar HS codes using a fine-tuned Sentence-BERT (SBERT) model [20]. It leverages MCD during inference to generate uncertainty estimates and improve ranking robustness using a hybrid approach.

Table 1: Corpus Statistics Across Train and Test Splits

Split	Descriptions	Unique HS Codes	Coverage (%)*	Avg. tokens
Train	171,247	4,924	71.4	158
Test	37,794	3,470	50.3	169
Reduced Test	1,000	598	8.7	172

* Coverage calculated against 6,895 total HS codes.

3.1 Data Collection

To ensure real-world adaptability, cross-domain generalizability, and temporal robustness, we utilized three diverse data sources totaling approximately 208,000 product-to-HS code mappings: United States Customs and Border Protection Customs Rulings (CBP CROSS) [28] from 1989 to present; European Binding Tariff Information (EBTI) [9] from 2004 to 2025; and U.S. Trade Representative (USTR) [19] rulings on machinery imports from 2024. After preprocessing and a randomized 80-20 train-test split, the final training dataset comprises 171,247 product rulings encompassing 4,924 unique HS codes, with varying description lengths, domain coverage, and linguistic attributes, as shown in Table 1.

3.2 Semantic Search Modeling

Based on testing from several base models (e.g., all-MiniLM-L6-v2 (MiniLM)[20], all-mpnet-base-v2 (MPNet)[23]), the best model, MiniLM, was selected based on its speed and strong performance. We fine-tune this base model using MNRL with hierarchically-stratified hard negatives to generate embeddings i.e., prioritizing codes from the same 4-digit heading, then 2-digit chapter, and defaulting to random codes only when hierarchical options are exhausted, to preserve the HS hierarchy while creating meaningful distinctions. This fine-tuning method [4] was chosen for its interpretability, ease of implementation, and its strong performance against other loss functions (e.g., contrastive triplet network loss function). Retrieval is performed by computing the cosine similarity score between the product description and HS code and returning the top-k ($k = 10$).

3.3 Uncertainty Quantification

Due to complexities in HS code structures, suggesting multiple options to the user along with uncertainty measurements can assist in making appropriate code assignment decisions. Therefore, we employ MCD, a technique more commonly used in neural networks, to provide confidence measures and rerank predictions based on stability across multiple stochastic forward passes.

For each product description p, we perform $M = 50$ stochastic forward passes with dropout enabled during inference. Each pass produces a ranked list of HS codes with associated cosine similarity scores, resulting in M different predictions.

To measure prediction stability and rerank results, we aggregate cosine similarity across all M samples. For each unique HS code c that appears in any of the samples, we compute two metrics:

- Average cosine similarity, measuring how strongly code c matches to the query:

$$\text{avg\_cos}(c) = \frac{1}{N_c} \sum_{m=1}^{M} \sum_{k=1}^{K} \cos(\theta_{m,k}) \cdot \mathbf{1}\{c_m^{(k)} = c\} \tag{1}$$

  where N_c is the total number of times code c appears across all samples and ranks, $\cos(\theta_{m,k})$ is the cosine similarity at rank k in sample m, and $c_m^{(k)}$ is the code at rank k in sample m.
- Top-3 frequency, measuring how consistently code c appears in the top-3 ranks:

$$\text{top3\_freq}(c) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{c \in \{c_m^{(1)}, c_m^{(2)}, c_m^{(3)}\}\} \tag{2}$$

  This metric ranges represents the proportion of samples where c appeared in the top-3 ranks. We focus on top-3 ranks as prior search studies show users primarily examine the top 1 to 3 ranked results when making decisions [12].

We empirically hypertuned the weight parameters on a validation set, testing combinations in increments of 0.1 and dropout samples $M = 20, 50, 100$. The configuration of $M = 50$ with weights 0.8 for average cosine similarity and 0.2 for top-3 frequency yielded optimal Recall@10 performance. This weighting prioritizes overall similarity strength while still accounting for rank consistency, based on the analysis that codes appearing frequently in top ranks with high similarity scores are more reliable predictions.

These measures are combined into the following stability score:

$$f_s(c) = 0.8 \cdot \text{avg\_cos}(c) + 0.2 \cdot \text{top3\_freq}(c) \tag{3}$$

The final HS codes search results are reranked in descending order of $f_s(c)$. This approach promotes predictions that consistently appear with high similarity scores across multiple stochastic samples, providing more reliable rankings than single-pass inference and interpretable confidence measures.

4 Experimental Results

We demonstrate the utility of Text2Trade using real product queries obtained from the United States Geological Survey of Mineral Commodity Summaries (USGS) [29] and out-of-sample CBP rulings. The queries from this testing phase are diverse in domain, query length (i.e., USGS has 1-2 word queries while CBP rulings are a few sentences), and year. To test the performance of the fine-tuned SBERT model, we evaluate performance using standard ranking metrics: Recall@k (k=1, 3, 5, 10), Normalized Discounted Cumulative Gain (nDCG)@10, and Mean Average Precision (MAP)@10. Recall@k measures the proportion of queries for which the true HS code appears within the top-k predictions. The nDCG@10 and MAP@10 assess ranking quality by rewarding correct codes appearing higher in the list and how well the system prioritizes relevant codes across the top-10 predictions, respectively.

4.1 Baseline Comparison

To demonstrate the effectiveness of Text2Trade, we compare against six baseline methods: Linear Support Vector Machine (SVM) [6], which achieved significant results in prior work on HS code classification [1]; BM25 [22] as a retrieval-based baseline; LSTM [11]

Table 2: Baseline Performance comparison

Model	R@1	R@3	R@5	R@10	nDCG @10	MAP @10	Time (ms)**
Complete Test set							
Text2Trade (no MCD)	**.740**	**.910**	**.940**	**.970**	**.851**	**.774**	**13.30**
LSTM	.221	.370	.436	.523	.362	.291	44.09
MiniLM*	.108	.220	.290	.395	.235	.171	8.72
MPNet*	.100	.194	.244	.322	.199	.148	13.55
BM25	.544	.727	.786	.838	.689	.612	78.20
Linear SVM	.567	.724	.749	.765	.676	.605	7.21
Reduced Test set							
Text2Trade (no MCD)	**.706**	**.872**	**.919**	**.957**	**.832**	**.766**	**13.30**
LSTM	.181	.343	.404	.498	.329	.259	44.09
MiniLM*	.104	.193	.268	.382	.221	.159	8.72
MPNet*	.100	.193	.238	.309	.194	.147	13.55
BM25	.521	.710	.770	.825	.673	.597	78.20
Linear SVM	.534	.720	.747	.762	.660	.590	7.21
Llama-4	.286	.453	.514	.589	.431	.365	2,680

* Pre-trained SBERT model

** Inference time per query

as a neural sequence model; two pretrained SBERT [20] to isolate the contribution of our fine-tuning strategy; and Llama-4-Maverick [27] as a representative large language model (LLM) using zero-shot prompting without fine-tuning.

We evaluate all non-LLM baselines on the complete test set (37,794 queries). Due to computational constraints, we additionally evaluate the LLM-based approach on a stratified subset of 1,000 descriptions, sampled proportionally from each HS revision year. This ensures that the reduced test set maintains the consistent distribution as the complete test set, ensuring representative evaluation. Table 1 summarizes the statistics of the full and reduced test sets.

As summarized in Table 2, Text2Trade significantly outperforms all baseline models in retrieval accuracy, with Recall@1 and Recall@10 of 74% and 97%, respectively. Its ranking quality is also superior, with nDCG@10 and MAP@10 scores that exceed all baselines. This performance holds on the complete test set, where Text2Trade continues to consistently achieve the best overall performance.

In addition to accuracy, Text2Trade demonstrates competitive search efficiency. As shown in Table 2, it takes approximately 13.6 milliseconds per query, making it three times faster than LSTM and six times faster than a traditional IR model such as BM25, which scores the entire corpus per query. Text2Trade's inference time is comparable to pretrained SBERT models while achieving significantly higher accuracy, demonstrating an efficient balance between speed and retrieval quality for real-world use.

4.2 Ablation Study

The effectiveness of MCD and MRNL is validated through ablation experiments on the reduced test set. Table 3 summarizes the results when each method is applied individually and in combination. The addition of MCD to a pre-trained SBERT model nearly doubles Recall@1. Fine-tuning SBERT with MNRL furthers the performance, increasing Recall@1 by almost seven times. Combining MCD with MNRL (Text2Trade) achieves the best performance overall. The fine-tuned approach with MCD reranking improves accuracy by 74% compared to pre-trained SBERT, demonstrating the superior HS code search capability of Text2Trade.

Table 3: Ablation study on stratified subset.

Component	R@1	R@3	R@5	R@10
Pretrained SBERT	.100	.193	.238	.309
+ MCD only	.199	.271	.326	.443
+ MNRL only	.706	.872	.919	.957
Fine-tuned + MNRL + MCD	**.743**	**.875**	**.929**	**.978**

4.3 Generalizability Analysis

We evaluate the domain generalizability of Text2Trade by evaluating the performance across three distinct and particularly challenging HS sectors: textiles, machinery, and chemicals and pharmaceuticals [31] due to granular descriptions of material composition, functionality, and technical specifications. Product descriptions from these sectors were selected from the reduced test set to assess whether performance holds consistently across domains with varying classification complexity.

As illustrated in Figure 3, Text2Trade substantially surpasses all baseline models, achieving Recall@10 of 78% in chemicals and pharmaceuticals, 89% in machinery, and 96% in textiles.

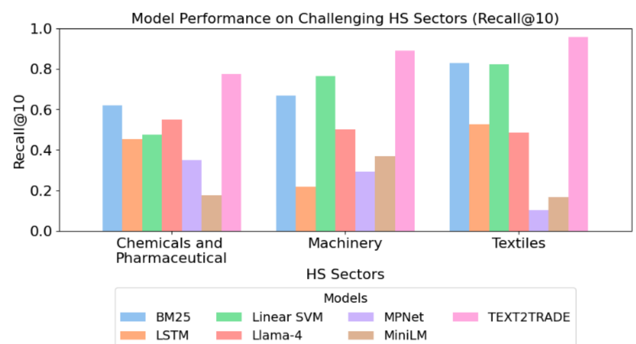

Figure 3: Comparison of baseline and Text2Trade performance on three challenging HS sectors using Recall@10.

5 Conclusion

In this paper, we presented Text2Trade, a semantic search system for HS code search that integrates MNRL-based fine-tuning and MCD as a reranking and uncertainty quantification method. Through empirical evaluation on diverse, publicly available trade data, we demonstrate significant improvements over traditional ML, keyword-based, and LLM baselines, with strong performance across challenging domains, while supporting human-in-the-loop workflows. These benefits have the potential to substantially reduce expert search time and mitigate financial losses from HS code misclassifications. Future work will include experimental studies to quantify time efficiency and financial gains.

Our approach is constrained by limited coverage of publicly available trade data (71.4% of HS codes). Future work with expanded datasets and targeted resampling could improve generalization on underrepresented codes. Additionally, this approach could be extended to granular tasks, such as 8- or 10-digit HS codes, or adapted to other hierarchical coding systems in domains like medical diagnosis codes, where similar semantic search challenges exist.

References

[1] Fatma Altaheri and Khaled Shaalan. 2020. Exploring Machine Learning Models to Predict Harmonized System Code. In *Information Systems*, Marinos Themistocleous and Maria Papadaki (Eds.). Springer International Publishing, Cham, 291–303. doi:10.1007/978-3-030-44322-1_22

[2] Pierre Baldi and Peter Sadowski. 2014. The dropout learning algorithm. *Artificial Intelligence* 210 (May 2014), 78–122. doi:10.1016/j.artint.2014.02.004

[3] US Census Bureau. [n. d.]. Schedule B. https://www.census.gov/foreign-trade/schedules/b/index.html

[4] Johnny Chang and Kaushal Alate. [n. d.]. "That was smooth": Exploration of S-BERT with Multiple Negatives Ranking Loss and Smoothness-Inducing Regularization. ([n. d.]).

[5] Xi Chen, Stefano Bromuri, and Marko Van Eekelen. 2021. Neural Machine Translation for Harmonized System Codes prediction. In *2021 6th International Conference on Machine Learning Technologies*. ACM, Jeju Island Republic of Korea, 158–163. doi:10.1145/3468891.3468915

[6] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

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

[8] Shaohua Du, Zhihao Wu, Huaiyu Wan, and YouFang Lin. 2021. HScodeNet: Combining Hierarchical Sequential and Global Spatial Information of Text for Commodity HS Code Classification. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 676–689. doi:10.1007/978-3-030-75765-6_54

[9] European Commission, Directorate-General for Taxation and Customs Union. 2025. European Binding Tariff Information (EBTI) BTI Consultation. https://ec.europa.eu/taxation_customs/dds2/ebti/ebti_consultation.jsp?Lang=en Place: Brussels, Belgium Published: Online Database.

[10] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, 1050–1059. https://proceedings.mlr.press/v48/gal16.html ISSN: 1938-7228.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, Salvador, Brazil, 154–161. doi:10.1145/1076034.1076063

[13] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports* 7, 1 (Dec. 2017), 17816. doi:10.1038/s41598-017-17876-z Publisher: Nature Publishing Group.

[14] Mengjie Liao, Lei Huang, Jian Zhang, Luona Song, and Bo Li. 2024. Enhanced HS Code Classification for Import and Export Goods via Multiscale Attention and ERNIE-BiLSTM. *Applied Sciences* 14, 22 (Jan. 2024), 10267. doi:10.3390/app142210267 Publisher: Multidisciplinary Digital Publishing Institute.

[15] Sungjun Lim, Kangjun Noh, Youngjun Choi, Heeyoung Lee, and Kyungwoo Song. 2025. Uncertainty-driven Embedding Convolution. doi:10.48550/arXiv.2507.20718 arXiv:2507.20718 [cs].

[16] Andrea Littler. 2022. TariffTel - The high price of using incorrect tariff codes. https://www.tarifftel.com/blog/the-high-price-of-using-incorrect-tariff-codes/

[17] Yuchen Lu, Ying Jin, and Xi Chen. 2023. Chapter 15: Recombination-based two-stage out-of-distribution detection method for traffic flow pattern analysis. https://www.elgaronline.com/edcollchap/book/9781803929545/book-part-9781803929545-26.xml Section: Handbook on Artificial Intelligence and Transport.

[18] Liangping Ma and John Kaewell. 2020. Fast Monte Carlo Dropout and Error Correction for Radio Transmitter Classification. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–5. doi:10.1109/WIFS49906.2020.9360887 ISSN: 2157-4774.

[19] Office of the United States Trade Representative. 2024. Temporary Exclusions for Machinery Used in Domestic Manufacturing. https://comments.ustr.gov/s/docket?docketNumber=USTR-2024-0020 Place: Washington, DC Published: Public Comments Docket.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. doi:10.48550/arXiv.1908.10084 arXiv:1908.10084 [cs].

[21] Cailen Gordon Robertson. [n. d.]. DSpace. https://research-repository.griffith.edu.au/items/98041ee4-4f38-400f-b78c-c271bdb0b133

[22] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval*, Vol. 3. Now Publishers, 333–389.

[23] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*. 16857–16867.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[25] Sédrick Stassin, Otmane Amel, Sidi Ahmed Mahmoudi, and Xavier Siebert. 2023. Similarity versus Supervision: Best Approaches for HS Code Prediction. In *ESANN 2023 proceesdings*. Ciaco - i6doc.com, Bruges (Belgium) and online, 175–180. doi:10.14428/esann/2023.ES2023-163

[26] Catalin Stoean, Ruxandra Stoean, Miguel Atencia, Moloud Abdar, Luis Velázquez-Pérez, Abbas Khosravi, Saeid Nahavandi, U. Rajendra Acharya, and Gonzalo Joya. 2020. Automated Detection of Presymptomatic Conditions in Spinocerebellar Ataxia Type 2 Using Monte Carlo Dropout and Deep Neural Network Techniques with Electrooculogram Signals. *Sensors* 20, 11 (2020). doi:10.3390/s20113032

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. doi:10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs].

[28] U.S. Customs and Border Protection. 2024. Customs Rulings Online Search System (CROSS). https://rulings.cbp.gov/ Place: Washington, DC Published: Online Database.

[29] U.S. Geological Survey. 2024. *Mineral Commodity Summaries 2024*. Technical Report. U.S. Department of the Interior, U.S. Geological Survey, McLean, Virginia. doi:10.3133/mcs2024

[30] Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty Estimation and Reduction of Pre-trained Models for Text Regression. *Transactions of the Association for Computational Linguistics* 10 (June 2022), 680–696. doi:10.1162/tacl_a_00483

[31] World Customs Organization. 2019. Revitalizing the Harmonized System: What is Needed for a 21st Century Harmonized System? WCO Conference on the Future Direction of the HS. 2–3 May 2019.

[32] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. 2023. Dirichlet-based Uncertainty Calibration for Active Domain Adaptation. doi:10.48550/arXiv.2302.13824 arXiv:2302.13824 [cs].