Automatic Product Classification in International Trade: Machine Learning and Large Language Models*

Ignacio Marra de Artiñano
Université Libre de Bruxelles

Franco Riottini Depetris
Inter-American Development Bank and UdeSA

Christian Volpe Martincus
Inter-American Development Bank and CESifo

This version: July 2023†

Abstract

Accurately classifying products is essential in international trade. Virtually all countries categorize products into tariff lines using the Harmonized System (HS) nomenclature for both statistical and duty collection purposes. In this paper, we apply and assess several different algorithms to automatically classify products based on text descriptions. To do so, we use agricultural product descriptions from several public agencies, including customs authorities and the United States Department of Agriculture (USDA). We find that while traditional machine learning (ML) models tend to perform well within the dataset in which they were trained, their precision drops dramatically when implemented outside of it. In contrast, large language models (LLMs) such as GPT 3.5 show a consistently good performance across all datasets, with accuracy rates ranging between 60% and 90% depending on HS aggregation levels. Our analysis highlights the valuable role that artificial intelligence (AI) can play in facilitating product classification at scale and, more generally, in enhancing the categorization of unstructured data.

Keywords: Product Classification, Machine Learning, Large Language Models, Trade
JEL Codes: F10, C55, C81, C88

*We would like to thank Peter Schott for valuable comments and suggestions. The views and interpretations in this paper are strictly those of the authors and should not be attributed to the Inter-American Development Bank, its executive directors, or its member countries.

†The latest version can be downloaded here.

Contact: Christian Volpe Martincus, Inter-American Development Bank, e-mail: christianv@iadb.org.

1 Introduction

Accurately classifying products is essential in international trade. Virtually all countries use the Harmonized System (HS) nomenclature to categorize products into tariff lines for both statistical and duty collection purposes. Misclassification, both intentional and unintentional, can be very costly. It can result in imprecise measurement of trade flows, inappropriate determination of origin, foregone duty collection, inadequate application of restrictions or prohibitions, significant delays to border monitoring and processing times, and the design and implementation of misguided trade policies, particularly trade remedies (such as countervailing duties, antidumping, and safeguards).

Traditionally, the bulk of product categorization tasks has been carried out manually, frequently based on experts' judgments, and has accordingly been extremely time-consuming.[1] As a consequence, classification is challenging for governments, firms, and researchers, especially on a large scale. Thus, the rise of cross-border e-commerce requires customs agencies to process several million small shipments per year. In many developing countries, this has generally resulted in most shipments being classified based on their value or size instead of the specific goods they consist of, which limits their ability to conduct risk assessments properly and that of their countries to accurately measure the composition of a growing portion of their international trade. Firms, in turn, particularly those that are small or have no previous experience in international trade, typically find it difficult to assign their products to HS codes and need to rely on costly specialized services to do so.[2] Last but certainly not least, various databases that could potentially provide inputs for novel, policy-relevant research report valuable product-level information according to product names or descriptions. This makes it hard for researchers to combine them, leading to imperfect merges with standard trade databases based on the HS nomenclatures. In this paper, we look at an important

[1] As highlighted by public customs' agencies resolutions, classification is often the object of firms' ex-ante consultations and is subject to ex-post adjustments.

[2] Furthermore, in an effort to reduce the wrong attribution of tariff lines, custom agencies often impose heavy misreporting fines, which can be burdensome for exporters.

example of one such attempt to classify products into the HS nomenclature using product descriptions.

The advent of machine learning (ML) is likely to reduce these classification efforts and increase their accuracy (see WCO, 2022a).[3] While there is an incipient literature that aims to assess the accuracy of ML for product classification, most existing models rely on tests on the same dataset used to train them. As a consequence, there is very limited evidence on how these models perform on real external datasets and hence on their general applicability. Further, such evidence is missing altogether in the case of large language models (LLM) such as GPT-3.5, which are yet to be tested at scale for this purpose.

In this paper, we will examine the performance of a variety of ML models, including GPT-3.5, at classifying products according to the HS nomenclature at different aggregation levels. In doing so, we will go beyond the train-and-test dataset and thus explicitly assess the external validity of the models. For this, we will use three different datasets: (i) a dataset containing product descriptions from the Chilean customs agency to train and test ML algorithms, following earlier literature; (ii) a dataset containing product descriptions from the customs agency of a different country, Paraguay; and (iii) a database of product descriptions from the United States Department of Agriculture (USDA).[4] This third data source describes products for which firms obtain an organic certification. In all cases, our analysis will be limited to animal, vegetable, and food products, since these are the product categories for which firms can obtain organic certification (Marra de Artiñano et al., 2023)[5].

Our analysis reveals that while standard ML algorithms performed very well within the test set, their accuracy dropped dramatically when these models were applied to datasets on which they were not explicitly trained. In contrast, GPT-3.5 performed very evenly across all datasets. Its accuracy was relatively high: it achieved percentages of approximately

[3]The BACUDA project run by the World Customs Organization (WCO) is an example of ongoing work using these techniques for customs applications.

[4]We use Paraguayan customs data because, like Chilean data, they are freely available

[5]This project was originally conceived with the objective to match product descriptions of organic certified firms with the HS Codes. In a future study we will approach the problem from a holistic point of view, i.e. for all products under HS classification.

60%—70% at the HS 6-digit level (highly granular product nomenclature), 70%—80% at the HS 4-digit level, and 80%—90% at the HS 2-digit level.[6].

There are several important applications for this sort of scalable automatic product classification that uses product descriptions as inputs. First, it could help customs agencies identify patterns of intentional or fraudulent product miscategorization. Second, it would make it easier for both policymakers and researchers to categorize product descriptions from unstructured data sources (such as those obtained from e-commerce transactions) using established product nomenclatures. Finally, it could be used to develop chatbots that give HS code suggestions from simple text descriptions, which would greatly facilitate tariff line attribution for firms engaged in international trade and even consumers participating in cross-border e-commerce.

To the best of our knowledge, this study is the first to apply GPT to the WCO's HS product classification and, more generally, to a large multiclass classification problem in economics. [7].

A number of previous studies have proposed alternative approaches to automatically classify products into HS codes across a large number of tariff lines. Spichakova and Haav (2020) use ML methods to provide 6-digit HS code predictions and recommendations using a model trained with product descriptions from the United States Bill of Lading 2017 database. They show that the algorithm achieves a hit rate of 80% on the test dataset. Ruder (2020) uses a variety of ML and deep learning models to classify product descriptions from the US Bill of Lading and reaches accuracy levels of approximately 60%. Chen et al. (2021) apply unsupervised ML and an off-the-shelf embedding encoder to automatically assess whether reported HS codes in cross-border import declarations are correct. They achieve an overall success rate of 71% on an HS 6-digit dataset provided by Dutch customs Turhan et al. (2015) adopt a different strategy whereby they use visual properties along with product labels and descrip-

[6]We also tested the performance of GPT 3.5 in mapping sector descriptions onto the North American Industry Classification System (NAICS). To do so, we used sectors reported by firms when registering with the online business platform *ConnectAmericas*. The results indicate that the GPT-3.5 model achieves an efficiency of more than 60% at the 6-digit level. These results are available from the authors upon request.

[7]Kocoń et al. (2023) carries out a significantly simpler classification analysis using only a few categories.

tions. The accuracy level they achieve is above 80% with 4-digit HS codes from a database of 4,494 binding tariffs published by the European Union in 2014. These papers use a single dataset, which is split into training and testing samples. Unfortunately, this approach does not allow the accuracy of the models on external datasets to be tested. This limitation is crucial because tariff databases often have significantly different product descriptions and text formats. One exception in this regard is He et al. (2021), who use data gathered directly from firms to train their models, along with a second dataset of product descriptions from a third firm that was not in the test dataset. However, they focus on very few HS products (12 6-digit potential product classifications) and their exercise is accordingly much simpler than product categorization across the universe of potential tariff lines.

We contribute to this literature on automatic product classification by assessing the accuracy of different ML algorithms on both the test-train-split dataset and two additional datasets for a large set of products. Our results indicate a very large decrease in the accuracy of standard ML algorithms outside the dataset on which the models are trained.

There is also recent literature that aims to apply GPT and other LLM models to text- based data in the social sciences. Some recent papers that use GPT include S. Hansen et al. (2023), Lopez-Lira and Tang (2023), A. L. Hansen and Kazinnik (2023), K.-C. Yang and Menczer (2023) and Ko and Lee (2023)[8]. S. Hansen et al. (2023) compare the performance of a predecessor of GPT-3 to their own model, WHAM, and find that WHAM outperforms GPT-3 in terms of the error rate at the task of classifying whether a job posting allowed the possibility of remote work at least one day per week. The authors also discuss the potential gains of adopting modern natural language processing (NLP) methods for text classification in economic environments. They suggest that other prediction problems using text in economics might similarly benefit from a large training sample combined with sequence embedding models, such as GPT-3.

[8] An exhaustive analysis of the recent literature using ChatGPT (and its adjacent models) is beyond the scope of this paper. Nevertheless, it is worth mentioning papers such as Noy and Zhang (2023) on the effects on productivity, Biswas (2023) on its potential role in health, and Kasneci et al. (2023) on its potential impact on education.

Lopez-Lira and Tang (2023)examine the potential of ChatGPT in predicting stock market returns by using analysis and the classification of news with potential impact for firms. Their analysis suggests that, even though ChatGPT is not specifically trained for this task, it produces superior results in terms of predicting stock market returns than other traditional sentiment analysis methods commonly used in finance due to the comprehensiveness of the model. In a similar vein, Ko and Lee (2023) show that ChatGPT effectively helps improve portfolio management by selecting asset classes that statistically outperform random choices in diversification and returns.

A. L. Hansen and Kazinnik (2023) use GPT-3 and GPT-4 to decipher Fedspeak, the language used by the Federal Reserve to communicate monetary policy decisions. Their results suggest that these models obtain the lowest numerical errors, the highest accuracy rates, and the highest measure of agreement relative to human classification when compared to other pretrained linguistic models and dictionary-based approaches. Finally, K.-C. Yang and Menczer (2023) use ChatGPT to study the credibility of news and conclude that they are able to correctly evaluate news sources by rating them.

We add to these papers by showing the usefulness of LLMs for product classification in international trade. We find that while GPT-3.5 performs slightly worse than traditional ML algorithms on the test-train-split dataset, it significantly outperforms these models on external databases. The reason is that LLMs are able to go beyond the specific context of the training dataset and thus have much higher external validity. Unlike traditional ML algorithms, they also require no additional data-cleaning or preprocessing, making them much simpler to use.

The rest of this paper is structured as follows. Section 2 describes the different data sources used in our analysis. Section 3 explains the methodological approach. Section 4 discusses the results of the classification process for the different databases. Finally, Section 5 concludes with a brief discussion of our results.

2 Data

In this paper, we used three different datasets: a database of product descriptions from Chilean customs, a database of product descriptions from Paraguayan customs, and a database of organic product descriptions from the USDA. The first database (Chilean customs) was used to train and test the ML algorithms. The second database (Paraguayan customs) was employed to test the external validity of our models. Finally, the third database (USDA) was used to further test the models outside the context of customs product descriptions.

2.1 Train-Test-Split Dataset: Trade Transactions from the Chilean Customs

To generate and train the ML models that attempt to predict the HS nomenclator for a set of target products, we used the universe of Chilean export and import transactions between 2009 and 2021 as our train-and-test dataset. This comprehensive dataset contains more than 104 million observations, with granular information on trade transactions, including granular HS codes and detailed product descriptions. As is usual in the literature, we split this dataset into separate training and testing subsets. The training data set was used to develop and refine our models, while the test dataset was used to assess their performance and accuracy.

We focused our analysis on the products in HS chapters 1–22, which encompasses agricultural, animal, and food products. As mentioned above, our ultimate objective in this work was to accurately classify organic product descriptions into HS product nomenclatures, and thus we exclusively trained and tested in the categories these products are found in. To keep the computational load manageable, we randomly selected 1 million product descriptions in these HS chapters from the Chilean customs dataset. Following the standard practice in the ML literature, we used 70% of this sample for training purposes and the remaining 30% for testing purposes.

2.2 External Dataset 1: Trade Transactions from the Paraguayan Customs

To test our algorithms against a dataset outside the training set, we used a random sample of product descriptions from trade transactions recorded by Paraguayan customs. As before, we restricted the sample to agricultural, animal, and food products (HS chapters 1–22). Importantly, for this dataset, we not only had the product descriptions but also the HS codes assigned by firms, which enabled us to directly observe the accuracy of the HS codes provided by the different ML algorithms and by GPT-3.5.

2.3 External Dataset 2: USDA Organic Product Descriptions

Finally, we used information on products for which the USDA has issued organic certification to Latin American firms (see Marra de Artiñano et al. 2023). The original dataset comprises more than 26,000 product descriptions. These texts vary substantially in terms of how specific and clean they are (that is, whether they use clear, easy-to-understand wording and do not contain odd symbols and so on). Thus, these descriptions may be significantly shorter than those usually found in customs databases (e.g., "maize" or "mangoes"), and may be highly specific or scant (e.g., "concentrate soursop pulp" or "ungurahui"). Table A1 in the appendix shows selected descriptions for illustrative purposes.

3 Methodology

Classification algorithms play a vital role in a wide range of ML applications (Sarker, 2021).[9] Multiclass classification, a particularly challenging task, is one of the most widespread uses for classification algorithms. In this case, the objective is to categorize the data into three or more different and mutually exclusive categories (Aly, 2005), in such a way that what is sought is to train one or several models that can correctly assign a set of uncategorized data

[9]They have been used extensively in areas such as NLP (Otter et al., 2020), image recognition (Fujiyoshi et al., 2019; Lai, 2019), and sentiment analysis Mitra (2020), among others domains. In recent years, breakthroughs in NLP and text mining have propelled the adoption of these algorithms in real-world applications (Kowsari et al., 2019).

to the correct categories. Formally, given a training dataset of the form (x_i, y_i) where x_i is the ith input and y_i is the ith class label that belongs to the set {3, ..., N} we want to find a model H such that $H(x_i) = y_i$ for new, uncategorized data.

The process of automatic product classification using ML models consists of several steps. First, the train-and-test data (in our case, the product descriptions in trade transactions from Chilean customs) needs to be preprocessed, which involves preliminary cleaning of the data, splitting it, tokenizing it, and extracting features. Second, the data must be divided into the training and testing sets. Third, a series of different ML algorithms are applied to the training set.

After performing these steps, we also tested the models on two alternative external databases (product descriptions in trade transactions from Paraguayan customs and the USDA organic product database). We used OpenAI's GPT-3.5 API to classify the different products through direct prompts and benchmark its performance against that of the ML models.

Our analysis was entirely conducted using Jupyter notebooks and Python open-source libraries such as NLTK, scikit-learn, spaCy, AST, and other commonly used libraries, along with the OpenAI library to conduct the GPT prompt requests.

3.1 Data Processing

As mentioned above, the Chilean customs dataset covers 2009–2021, contains more than 104 million observations, and lists 12,934 different products at the HS 8-digit level. We processed this dataset by first restricting the product descriptions to those in chapters 1–22 of the HS schedule, which correspond to animal, vegetable, and food manufacturing products. This first filter reduced the total number of observations to approximately 12 million and the total number of unique 8-digit HS codes to 2,866.[10] We then proceeded to randomly select 1 million product descriptions in an effort to reduce the computational burden of the exercise.

[10] In addition, we filter out 469,435 observations that do not correspond to any known product according to the standard HS nomenclature (e.g., 16.00.00).

To preprocess the product descriptions, we performed a series of tasks that are summarized in table 1:

Table 1: Preprocessing of product descriptions

Step	Description
Text preparation	We imported the Natural Language Toolkit (NLTK) library and apply the "word tokenize" function to break the text into individual words (tokens). This was crucial, as it made postprocessing of text and feature extraction easier.
Lowercase	We converted all words to lowercase using a lowercase function. This helped to ensure that words are treated consistently in subsequent steps and to reduce data complexity.
Removal of non-ASCII characters	We applied a function to remove non-ASCII characters, except for the letter "ñ". This allows us to standardize and simplify the text, thus facilitating subsequent analysis.
Converting numbers written in words to digits	We used a function from the NLTK package to convert numbers written in words to digits. This helped reduce the complexity of the text and made it easier to extract relevant features.
Stop-word removal	We used a function to remove stop-words that do not provide relevant information for analysis, such as prepositions and conjunctions. This helped reduce the complexity of the text and allowed us to work on the most significant words.
Lemmatization	The lemmatize functions were used to transform words into their base or lemma form. This helped reduce the complexity of the text by grouping similar words together and made it easier to identify patterns in the data. [11]
Removing words that are not in English or Spanish	We applied a function to remove words that are not in English or Spanish. This helped focus the analysis on the relevant languages and reduced noise in the data.
English and Spanish noise removal	We applied some functions to remove irrelevant, noisy words in English and Spanish. This helped reduce noise in the data and allowed the most relevant words to be used for analysis.

Source: Authors' own elaboration.

By cleaning and preprocessing the text in the product descriptions as described in these steps, we got the data ready to be used properly with ML models and ensured that the models generated were accurate and efficient at estimating HS codes. Table A2 in the appendix illustrates the application of this procedure to a selected product description and shows the results thereof. This example provides a clear idea of the complexity of dealing with certain descriptions and demonstrates the importance of simplification if they are to be uses as inputs for traditional ML algorithms.

3.2 Traditional ML Algorithms

We used several different ML models for our multiclass classification problem. While offering an extensive explanation of such models is beyond the scope of this paper, this section contains a brief review of some of their characteristics, based primarily on Kowsari et al. (2019) and Aggarwal and Zhai (2012):

1. **Support Vector Machine (SVM)**: one of the most efficient ML algorithms since its introduction in the 1990s. SVM is a supervised learning algorithm that identifies the optimal hyperplane that separates data points into their respective classes and maximizes the margin between the classes. The key in this classifier is to "determine the optimal boundaries between the different classes and use them for the purposes of classification" (Aggarwal & Zhai, 2012).

2. **Rocchio**: a traditional and efficient method for text categorization. The algorithm represents documents as vectors in a high-dimensional space and calculates the centroid for each category. To classify a new product description, the algorithm measures the similarity of each to the centroids and assigns it to the closest category.

3. **Logistic Regression**: a linear model for binary classification, which can be extended to multiclass classification problems like categorizing product descriptions. Using a logistic function, the model estimates the probability of a product description belonging

to a specific class. The class with the highest probability is then assigned to the product description.

4. **k-Nearest Neighbors (k-NN)**: searches for the k most similar or closest items to the new object we want to classify, and then decides which category it belongs to, based on the most common category among its nearest neighbors.

5. **Random Forest**: an ensemble learning method that constructs multiple decision trees during training and combines their predictions to improve classification accuracy. This method can handle large datasets and effectively classify product descriptions into various HS chapters, despite the fact that it is quite slow to create predictions once trained.

6. **Naive Bayes**: a probabilistic classifier based on Bayes' theorem, which assumes independence between features. Although this assumption is often not valid in real-world applications, Naive Bayes classifiers still perform well in many cases. The classifier is particularly effective for text categorization tasks.

7. **Decision Tree**: a flowchart-like structure that can be used for classification tasks. The tree is built by recursively splitting the dataset based on the feature that provides the best separation into classes.

3.3 LLMs: GPT-3.5

GPT-3.5 is an advanced large-scale language, deep learning model.[12] It uses transformer architecture to understand and generate human-like text. With billions of parameters and the ability to learn from vast amounts of text data, it has been fine-tuned to excel in a wide range of NLP tasks.

Some of the notable properties of GPT-3.5include its autoregressive nature, which allows it to generate contextually relevant and coherent text by predicting the next word in a se-

[12]GPT-3.5 was developed by OpenAI. In our analysis, we use the GPT-3.5 version (internally called "gpt-3.5-turbo"), which powers the publicly available version of the ChatGPT chatbot. A more recent and powerful model, GPT-4, became available on March 14, 2023

quence given the previous words. The model is trained using unsupervised learning with a vast dataset that includes websites, books, and articles. Although the knowledge cut-off point for GPT-3.5 is September 2021, it is still be a powerful tool for various NLP tasks and can be adapted for specific use cases, such as assigning HS product nomenclature codes, in this instance.

We applied the model by asking it to assign an HS category based on the product description we provide. For that purpose, we gave it a system command to act as a wizard that assigns 6-digit HS codes and then asked it to do so for a given product description. In this regard, it is worth mentioning that we asked it not only to assign each product an HS code but also to provide its best estimate if the product description was not clear enough, thereby "forcing" it to make a guess.

Preparing datasets for use with the model (that is, the data processing described in section 3.1) was not essential. When working with LLMs, which are trained on a diverse range of text typologies, preprocessing data may not be needed and may even be disadvantageous as it might obscure valuable contextual information. We therefore merely input orders one at a time, allowing GPT-3.5 to categorize products individually. The prompt used and the completion request associated with it are presented in the appendix (section 3).

4 Classification Results

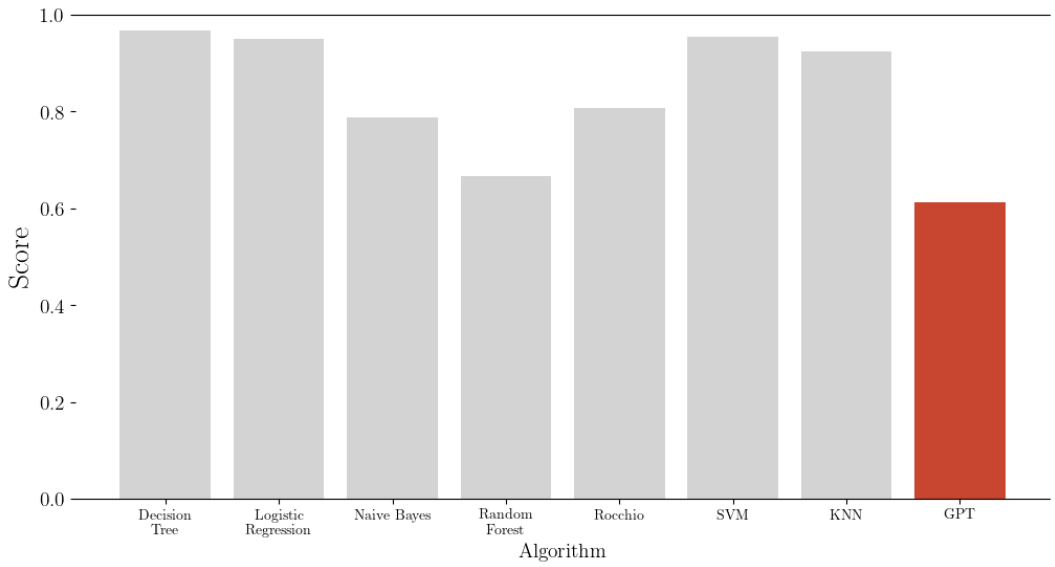
4.1 Results on the Train-Test-Split Dataset: Trade Transactions from the Chilean Customs

Figure 1 shows the accuracy of the different models on the Chilean customs data. It is important to stress that this is the dataset on which the ML algorithms are trained. Note that GPT-3.5 is not "trained" using any of the datasets, since the outcomes are obtained from direct prompts to the model through the API. The trained algorithms had very high accuracy levels on the test dataset, especially in the case of the Decision Tree, Logistic Regression, and SVM algorithms. The results of this test are typically used to assess the predictive capability

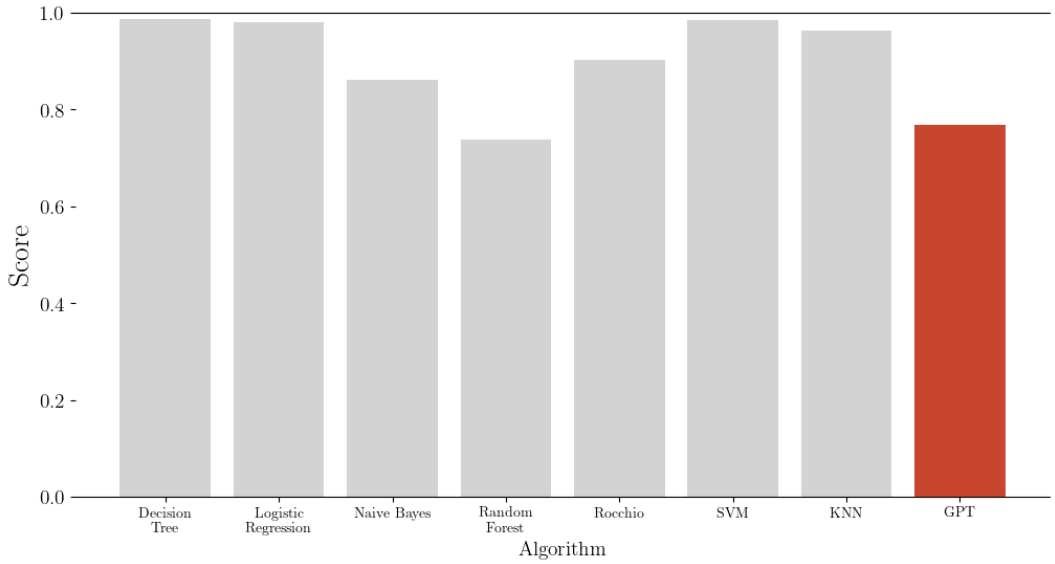
of an algorithm.

As expected, the accuracy levels were higher when less granular product categories were used (see figures 1b and 1c). However, this increase in success rates is uneven. For example, the hit rate of the GPT-3.5 model increased by 15 percentage points (from 62% to 77%) when moving from HS 6-digit codes to HS 4-digit codes and by an additional 9 percentage points (from 77% to 86%) when HS 2-digit codes were used. These findings indicate that GPT-3.5 predicts the broad category of products very well.

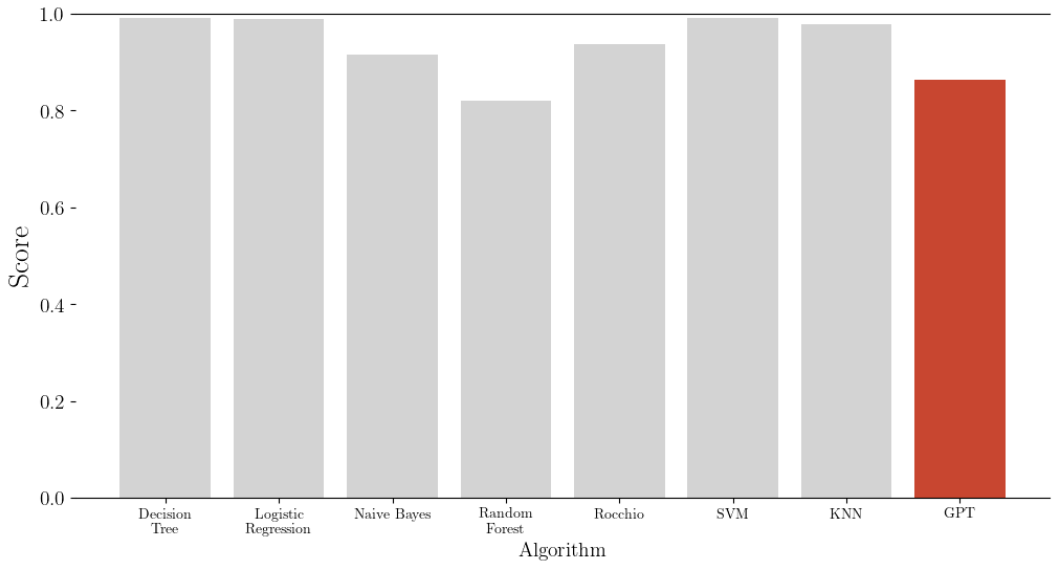
Figure 1: Algorithm's Accuracy in the Test-Train-Split Dataset: Chilean Customs.



(a) HS-6 digit level



(b) HS-4 digit level



(c) HS-2 digit level

Source: Authors' calculations based on Chilean customs data.

In this and the following subsection, we tested the ML algorithms outside the dataset on which they were trained. This was very important because the usefulness of such algorithms in real-world applications depends on their external validity. Real data imposes a clear chal-

lenge in this regard. It features a variety of product descriptions, including different formats. Hence, a model performing well on the training dataset may not be indicative of how well it will accomplish other classification tasks. To explore this, we compared the models using data that was not part of the test dataset. Specifically, we selected a random sample of 10,000 product descriptions from Paraguayan customs records. This allows for a fairer comparison of ML models and GPT-3.5, since it confronts both models with data on which neither was explicitly trained. The results are presented in figure 2.

Traditional ML algorithms did not perform well when tested using real-world data on which they were not trained. Their accuracy rates dropped below 30% for 6-digit HS codes. In contrast, the GPT-3.5 algorithm performed much better, correctly assigning around 60% of the product codes. These results are similar to those obtained on the Chilean dataset. This points to the consistency of GPT-3.5 in automatic product classification across customs datasets.[13]

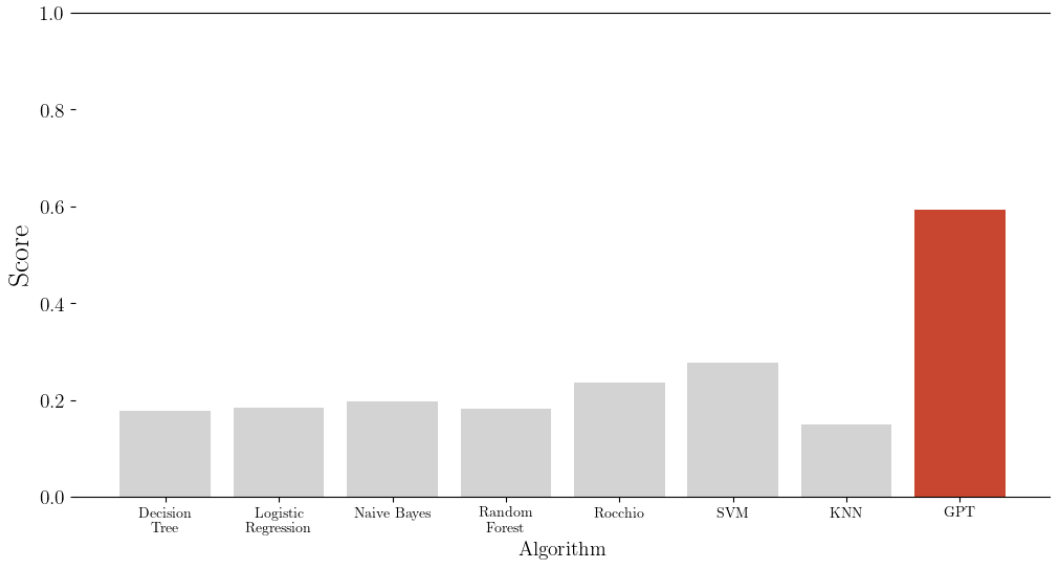
Next, we proceeded to check how the different algorithms performed at more aggregate levels. This allowed us to better understand how these algorithms work and where the highest rates of success/failure occur. Figure 2b shows the accuracy of the different algorithms when using 4-digit HS codes. Once again, conventional ML algorithms achieved a maximum accuracy level of 37%, while GPT-3.5 reached 77%, a 17-percentage-point increase in its hit rate compared with 6-digit codes.

Figure 2c reports the results for the more aggregated 2-digit classification. In this case, the GPT-3.5 algorithm achieved more than 90% accuracy. However, it should be noted that the performance of the conventional ML algorithms also improved significantly, with the Decision Tree reaching 73%. This indicates that all algorithms can predict the HS chapter that a product belongs to relatively well.[14]

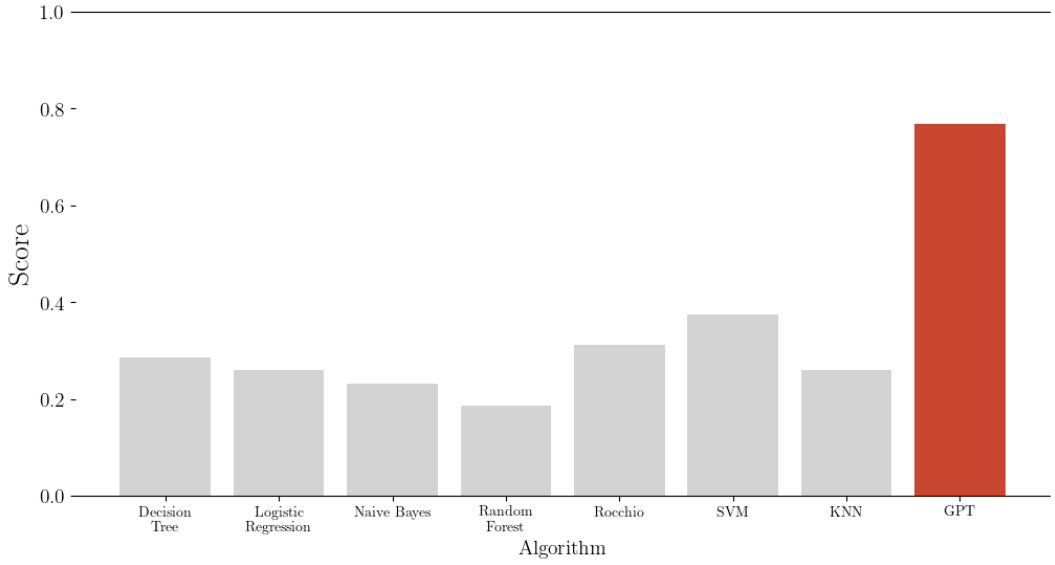
[13] In the appendix, we show GPT-3.5's accuracy at the HS 6-digit level for each broad HS 2-digit category. We failed to find any pattern, which suggests a high level of consistency in its average performance

[14] In addition, a comparison of Figures 2a, 2b and 2c reveals differences in terms of the best-performing conventional ML algorithm. While at the HS 6-digit level SVM has the highest accuracy rate, Decision Trees seem to outperform other methodologies at a less disaggregated level.

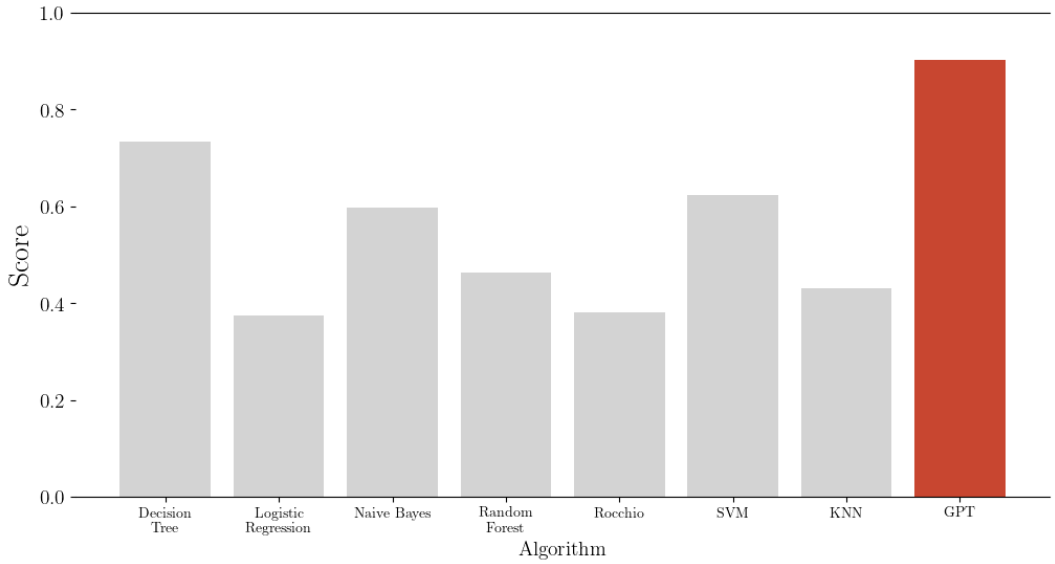
Figure 2: Algorithm's Accuracy in the First External Dataset: Paraguayan Customs.



(a) HS-6 digit level



(b) HS-4 digit level



(c) HS-2 digit level

Source: Authors' calculations based on Paraguayan customs data.

Despite the relative ease of predicting product chapters, GPT-3.5 performed consistently better than the other models. In the appendix, we show that GPT-3.5 performed well across all HS chapters included in our analysis (section 4).

4.2 Results on the External Dataset 2: USDA Organic Product Descriptions

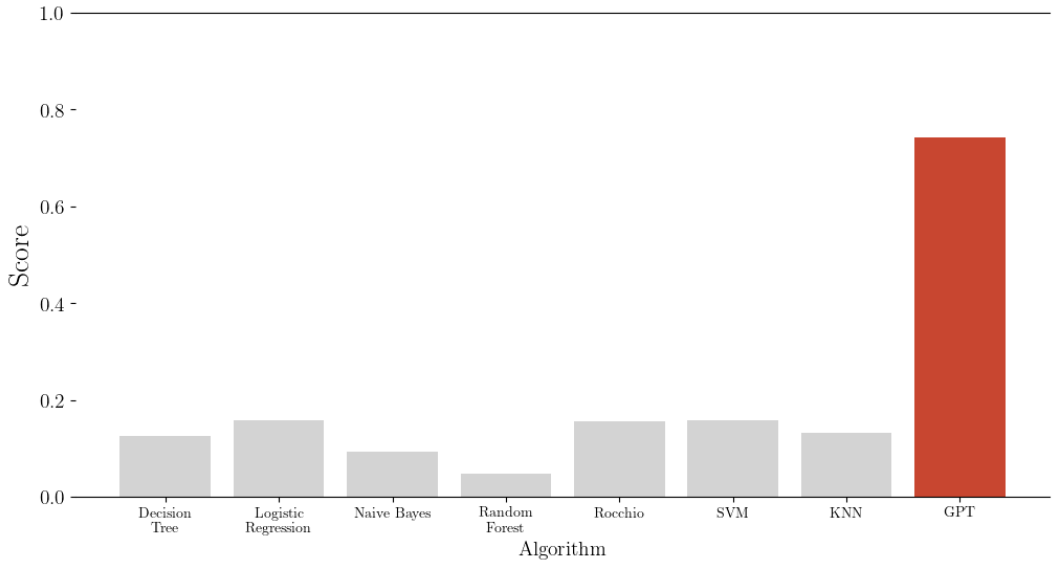
Finally, we assessed the ability of conventional ML algorithms and GPT-3.5 to accurately predict HS codes using text formats that differ from those traditionally used in customs. To do this, we used a set of descriptions of products for which Latin American firms are certified as organic producers and sellers by the USDA. As mentioned above, these product descriptions have different formats and vary significantly in terms of depth and specificity, which makes them potentially harder to categorize than the average customs product description. Furthermore, although this type of text contains descriptions of products, it does not specify the respective HS codes for each. Consequently, it cannot be used to train ML models to predict these. Similar cases can be found in many other data sources, such as cross-border e-commerce shipments, bank transactions, and survey-based descriptions.

To conduct this exercise, we selected a random sample of 1,000 descriptions of USDA certified organic products and classified these by hand into 6-digit HS tariff lines. The results are fully in line with those based on the Paraguayan customs external dataset: the standard ML algorithms performed significantly worse than GPT-3.5.[15]

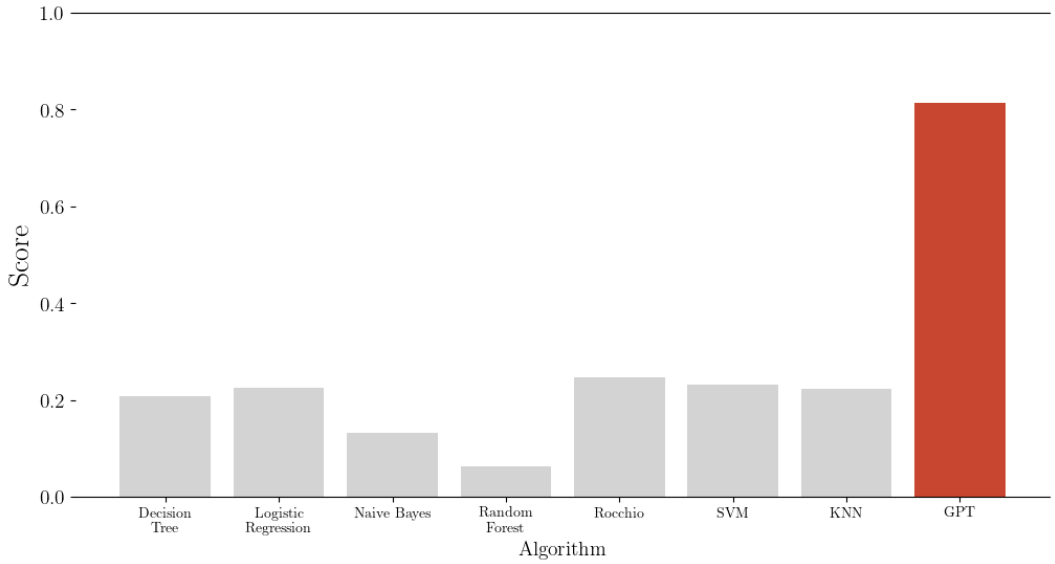
Figure 3a shows the accuracy at the HS 6-digit level. The GPT-3.5 model achieved a success rate of 74.1%, while the traditional ML models scored 15% at most (Rocchio model). The differences were similar when HS 4-digit codes were used: the accuracy of GPT-3.5 was over 80%, an improvement of 6 percentage points on the HS 6-digit level. Among the traditional ML algorithms, the maximum hit rate increased to 26% (again, the Rocchio model). Finally, at the HS 2-digit level, GPT-3.5 classified almost 88% of the product chapters correctly (i.e., a 7-percentage-point improvement on the 4-digit classification). It is noteworthy that this further widened the performance gap between GPT-3.5 and traditional ML models, whose success rate remained low even at these broader aggregation levels.

[15]To test the difference in performance from an increase of one order of magnitude in the number of products classified, we conducted a sensitivity analysis, in which we randomly divided the sample into 10 groups of 100 product descriptions and examined their accuracy. We found that GPT performed very similarly across the 10 groups, with a standard deviation of just 0.0136. We also did this for the other datasets (see appendix, section 5).

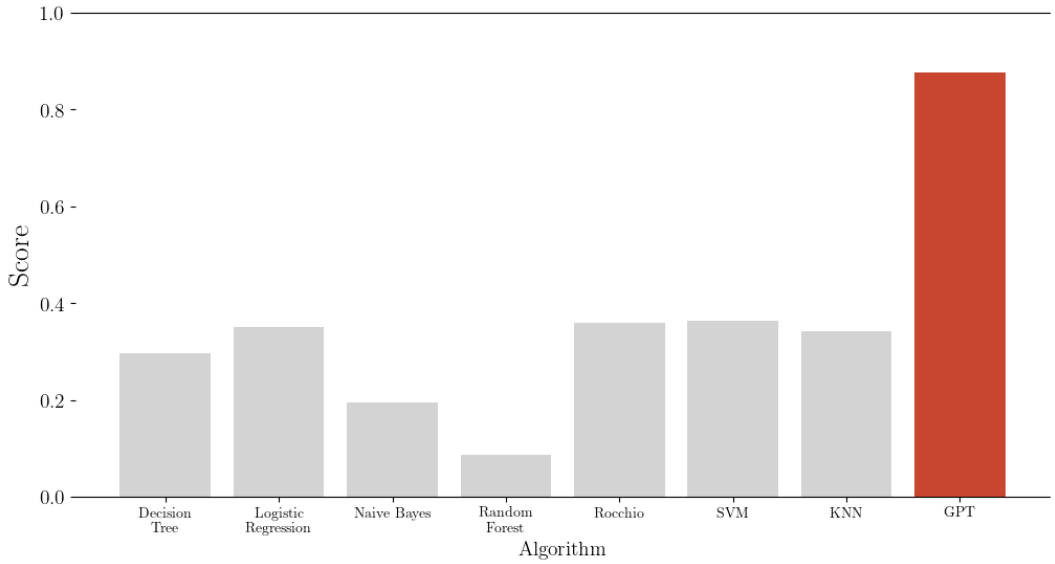
Figure 3: Algorithm's Accuracy in the Second External Dataset: USDA Organic Classification.



(a) HS-6 level



(b) HS-4 level



(c) HS-2 level

Source: Authors' calculations based on USDA data.

5 Discussion and Conclusions

The GPT-3.5 model showed high accuracy rates when classifying products according to the HS nomenclature. Traditional ML algorithms performed very well on their training dataset but their performance dropped dramatically when they were tested on external data. In such external validity tests, GPT-3.5 significantly outperformed these algorithms. Importantly, this was the case even when the ML models were trained with 1 million observations of high-quality customs product descriptions and then subsequently tested on high-quality descriptions from a different customs agency from the same region.

Another major advantage of GPT-3.5 is its ability to work with product descriptions in different languages. Throughout our analysis, we used data in English and Spanish, but GPT-3.5 is likely to perform very well across many other languages in which large amounts of data are publicly available (e.g., Chinese, French, German, etc.). Importantly, it is also able to successfully handle regional variants of the same language. One interesting example from our study was *Physalis peruviana*, a fruit, typically known as "goldenberry" in English.

Our Chilean training data refers to them as "uchuva," but the fruit goes by other names in different countries: "aguaymanto" in Peru, "uvilla" in Ecuador, and "fisalis" in Spain. ML algorithms trained on the Chilean data failed to identify these regional variations and thus misclassified the product, whereas GPT-3.5, trained on a much wider set of texts, recognized the fruit and classified it properly. This is an example of how the wide training dataset of LLMs allow them to outperform standard ML algorithms.

LLMs with chat interfaces are also significantly simpler since they do not require data-cleaning and preprocessing routines. Performing these tasks with traditional ML algorithms can be rather time-consuming and resource-intensive, especially those related to feature extraction. [16] While the API is necessary to work with GPT-3.5 at scale, the standard interface enables the classification functionality to be integrated easily into existing systems or applications. In our analysis, we worked with the base model, without making further adjustments, but GPT-3.5 could also be adapted for use with specific data through its fine-tuning mechanism.

LLMs can therefore be especially useful in comprehensive unilateral, regional, and multilateral trade policy initiatives involving product classifications over time and across countries (e.g., trade facilitation)

In terms of costs, the LLM we used (GPT-3.5) is relatively inexpensive, except at a very large scale.[17]. Importantly, open-source LLMs are becoming increasingly competitive, and we expect them to be able to perform at a high level in product classification tasks in the

[16] We also assessed the model against a manual classification carried out by a research assistant (RA) using a sample of 100 observations. The results indicate that, while the RA's accuracy was slightly above that of GPT-3.5 at the 6-digit level, the difference fades when more aggregate classification levels are considered. At the 2-digit HS level, GPT-3.5 performed slightly better than the RA. It is worth stressing that while the RA needed four hours to accomplish the task, GPT 3.5 completed it in just one minute. This suggests that there is potentially a tradeoff between accuracy and time for highly disaggregated classifications in small samples. The terms of this tradeoff are highly likely to change as the number of observations increases, with GPT-3.5 clearly emerging as the better approach for large samples, especially given that human working time increases at a nonlinear rate due to marginal decreasing returns.

[17] In our work, we used the latest, least expensive version of the GPT-3.5 model, "gpt-3.5-turbo". Without going into the billing system works in detail, our estimate is that the total cost of classifying a dataset of 10,000 standard customs product descriptions is approximately $3.20. (see the pricing)

short term.[18] Benchmarking automatic product classification across different LLMs (including open-source models) and different fine-tuning methods remains an important avenue for future research.

[18] See, for instance, Falcon and LLaMA.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163–222.

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, *19*, 1–9.

Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of Biomedical Engineering*, *51*(5), 868–869.

Chen, H., Van Rijnsoever, B., Molenhuis, M., van Dijk, D., Tan, Y. H., & Rukanova, B. (2021). The use of machine learning to identify the correctness of hs code for the customs import declarations. , 1–8.

Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS research*, *43*(4), 244–252.

Hansen, A. L., & Kazinnik, S. (2023). Can chatgpt decipher fedspeak? (Available at SSRN)

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). Remote work across jobs, companies, and space. (w31007).

He, M., Wang, X., Zou, C., Dai, B., & Jin, L. (2021). A commodity classification framework based on machine learning for analysis of trade declaration. *Symmetry*, *13*(6), 964.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Ko, H., & Lee, J. (2023). Can chatgpt improve investment decision? from a portfolio management perspective. *From a Portfolio Management Perspective*.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., … Kazienko, P. (2023). Chatgpt: Jack of all trades, master of none. *Information Fusion*, 101861. doi: 10.1016/j.inffus.2023.101861

Korinek, A. (2023). *Language models and cognitive automation for economic research* (Tech. Rep. No. w30957). National Bureau of Economic Research.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Lai, Y. (2019). A comparison of traditional machine learning and deep learning in image recognition. In *Journal of physics: Conference series* (Vol. 1314, p. 012148).

Lee, J., Choi, K., & Kim, G. (2021). Development of a natural language processing based deep learning model for automated hs code classification of the imported goods. *J Digit Contents Soc*, *22*(3), 501–508.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Lund, B. D., & Wang, T. (2023). Chatting about chatgpt: How may ai and gpt impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29.

Marra de Artiñano, I., Scattolo, G., Volpe Martincus, C., & Zavala, L. (2023). The value of organic certifications. *IDB Working Paper*. (Forthcoming)

Mitra, A. (2020). Sentiment analysis using machine learning approaches (lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, *2*(03), 145–152.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence.
(Available at SSRN 4375283)

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604–624.

Ruder, D. (2020). *Application of machine learning for automated hs-6 code assignment* (Unpublished master's thesis). University of Tartu, Institute of Computer Science.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Spichakova, M., & Haav, H. M. (2020). Application of machine learning for assessment of hs code correctness. *Baltic Journal of Modern Computing*, *8*(4), 698–718.

Turhan, B., Akar, G. B., Turhan, C., & Yukse, C. (2015). Visual and textual feature fusion for automatic customs tariff classification. , 76–81.

Xu, C.-J., & Li, X.-F. (2019). Research on the classification method of hs code products based on deep learning. *Mod. Comput.*, *01*, 13–21.

Yang, K., Ji, S., Zhang, T., Xie, Q., & Ananiadou, S. (2023). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Yang, K.-C., & Menczer, F. (2023). Large language models can rate news outlet credibility. *arXiv preprint arXiv:2304.00228*.

Appendix

1 Organic Descriptions

Table A1: Sample of 10 randomly chosen organic product descriptions

Original Product
Ungurahui (Oenocarpus Bataua)
soy beans
Plátanos/Bananos - 1 Traboar_Finca Genoveva (F)
Banana puree acidulated deep frozen
Organic aseptic concentrate soursop pulp
Organic white corn powder
Banana puree without seeds
Organic coco
Safflawer
Maca flour - pre cooked

Source: Own elaboration based on USDA

2 Preparation steps of descriptions.

Table A2: Preparation steps of a random selected description.

Step	Result
Initial description	FROZEN DOUGHS EUROPASTRY-F CODE-81299 BERLIDOTS BOMBOM FOOD PREPARATION BASED ON WHEAT FLOUR AND WATER IN BOXES OF 36 UNITS FOR HUMAN CONSUMPTION
Text preparation	['FROZEN', 'DOUGHS', 'EUROPASTRY-F', 'CODE-81299', 'BERLIDOTS', 'BOMBOM', 'FOOD', 'PREPARATION', 'BASED', 'ON', 'WHEAT', 'FLOUR', 'AND', 'WATER', 'IN', 'BOXES', 'OF', '36', 'UNITS', 'FOR', 'HUMAN', 'CONSUMPTION']
Lowercase	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Removal of non-ASCII characters	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Converting numbers written in words to digits	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Stop-word removal	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'wheat', 'flour', 'water', 'boxes', '36', 'units', 'human', 'consumption']
Lemmatization	['frozen', 'dough', 'europastery-f', 'code-81299', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']
Removing words that are not in English or Spanish	['frozen', 'dough', 'code', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']
English and Spanish noise removal	['frozen', 'dough', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']

Source: Authors' calculations based on Chilean customs data.

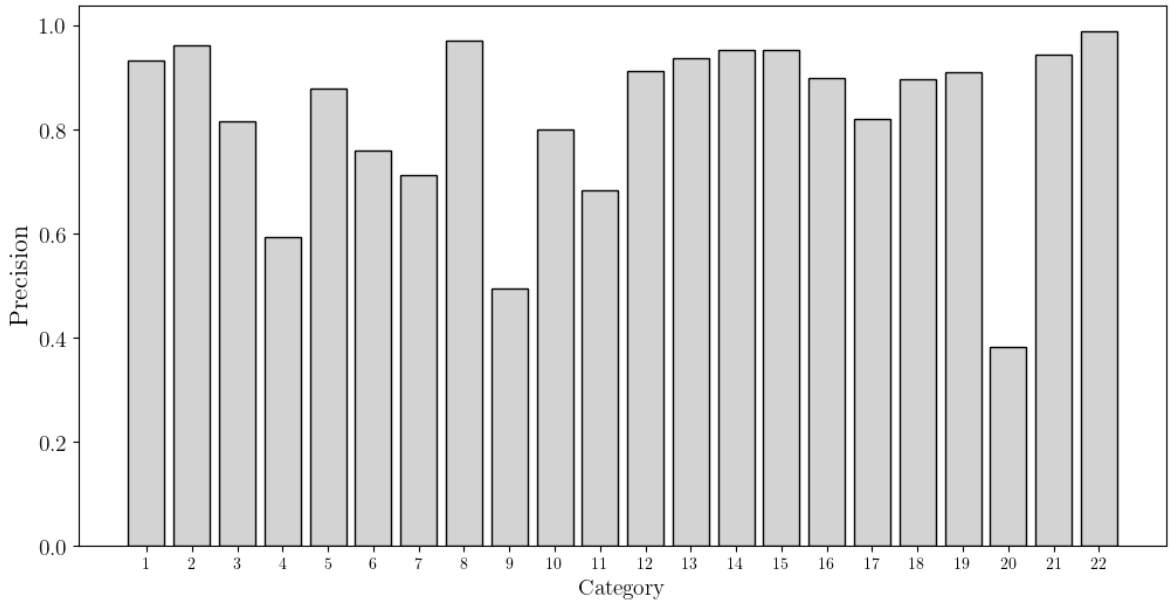
3 GPT-3.5 prompt.

```
@backoff.on_exception(backoff.expo, openai.error.RateLimitError, max_time=60)
def assign_code_forced(row, column):
    text = row[column]
    modelo = "gpt-3.5-turbo"
    try:
        response = openai.ChatCompletion.create(
            model=modelo,
            messages=[
                {"role": "system", "content": You are a helpful assistant that
                    assigns product codes in the HS6 product nomenclature
                    categorization."},
                {"role": "user", "content": f'Please assign the harmonized system
                    code number in the HS6 for the following description:"{texto}"
                    . Return "Code: number here". If you are unsure of the
                    classification, provide your best possible option'}],
            temperature=0.1)
        assigned_code = response['choices'][0]['message']['content']
        return assigned_code
    except json.JSONDecodeError:
        print(f"JSON Decode Error in text: {text}")
        return None
```

4 Accuracy in Specific HS Chapters

In this appendix, we show the results for the different product categories, showing where the GPT-3.5 model is most efficient. In both figure A4a and figure A4b, the data is shown up HS chapters 1–22, which are our target chapters, as we explained in the methodological section.

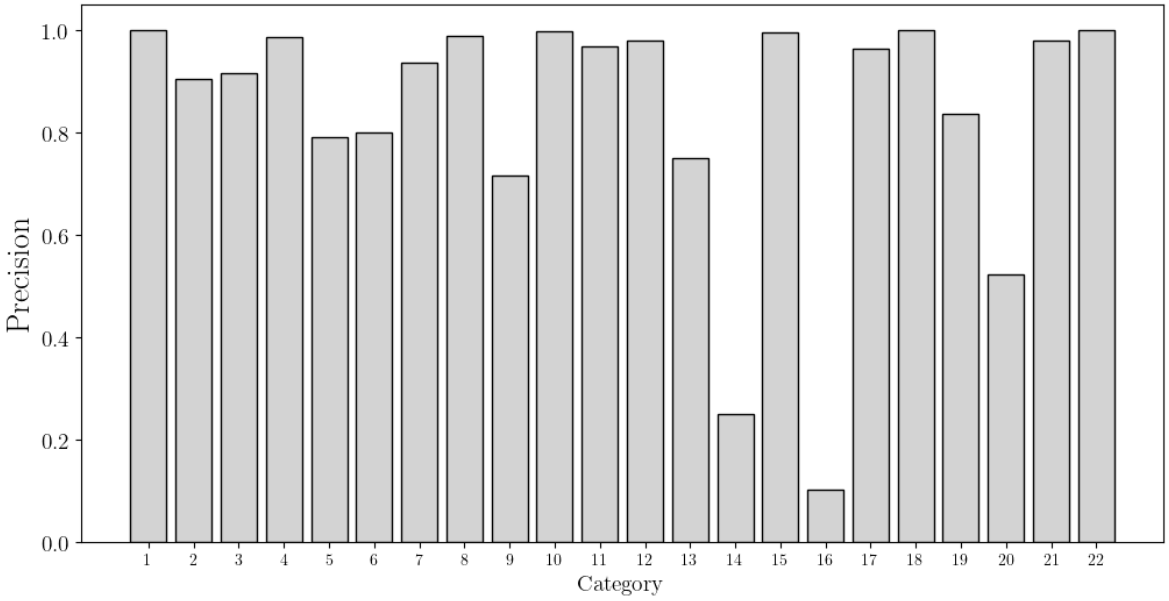
Figure A4a: Algorithm's accuracy in different HS chapters. Chilean dataset



Source: Authors' calculations based on Chilean customs data.

In figure A4a which shows the Chilean data we used in the training set for the algorithms, we see that the chapters with the lowest hit levels are 4, 9, and 20, which refer to "Dairy produce; birds' eggs; natural honey; edible products of animal origin," "Coffee, tea, mate and spices" and "Preparations of vegetables, fruit, nuts or other parts of plants," respectively. However, despite being relatively low in the chart, their accuracy scores are 0.70, 0.80, and 0.66, which are still good metrics.

Figure A4b: Algorithm's accuracy in different HS chapters. Paraguayan dataset



Source: Authors' calculations based on Paraguayan customs data.

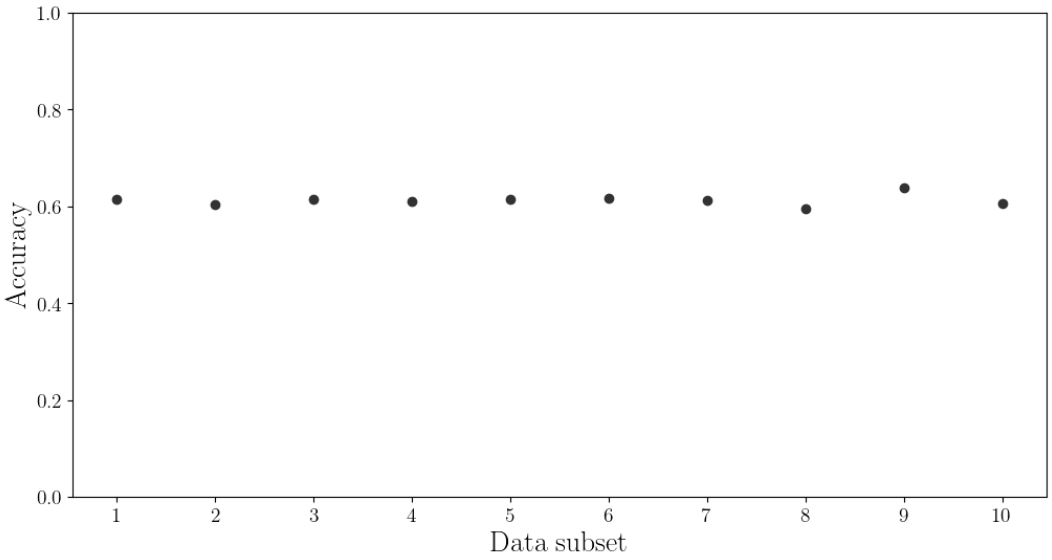
Using the Paraguayan data, we found that the efficiency was low in two categories: "Vegetable plaiting materials; vegetable products not elsewhere specified or included" (HS chapter 14), with 25% accuracy, and "Preparation of meat, of fish or of crustaceans, molluscs or other aquatic invertebrates" (HS chapter 16), with 10% accuracy. Other categories with relatively low accuracy levels are "Preparations of vegetables, fruit, nuts or other parts of plants" (HS chapter 20) and "Coffee, tea, mate and spices" (HS chapter 9).

A deep analysis of why the model fails most in these particular chapters goes beyond the scope of this paper. However, it may be strongly related to the quality of the data we are asking the algorithm to classify, that is, the product description shown by customs. The quality of the data and product descriptions might be significant factors affecting the model's performance, and it would be worth investigating this further. Improving data quality could lead to better results and greater accuracy across all chapters.

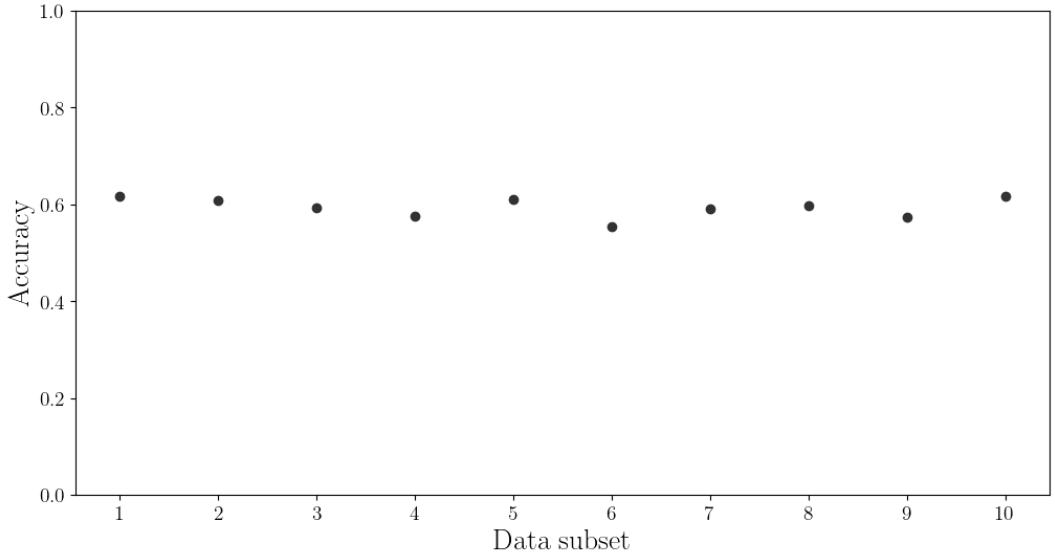
5 Subsample accuracy by reducing a degree of magnitude

In this appendix, we show the results of the efficiency point estimators by decreasing the order of magnitude of the sample, in order to show that the results are not affected by decreasing the number of observations classified using the algorithm. Figure A5a shows the point estimates after dividing the sample of observations from Chile into 10 groups. In this case, the observations have a standard deviation of 0.0034. Figure A5b does the same for the Paraguayan dataset. In this case, the standard deviation is very similar, at 0.0048. Finally, figure A5c does the same for the 1,000 classified observations from the USDA organic product database, for which the standard deviation is slightly larger, at 0.0136.

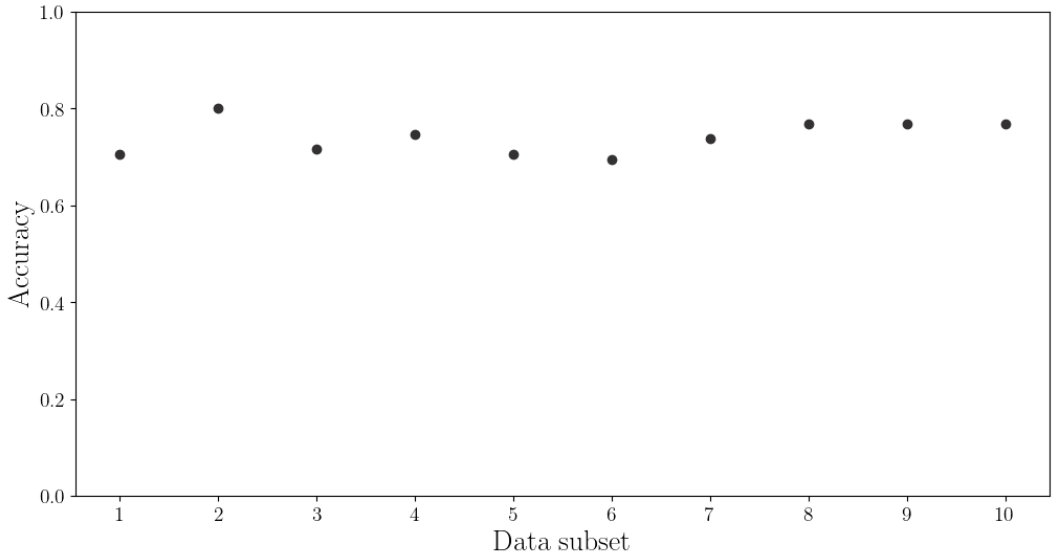
Figure A5: Algorithm's accuracy in the test dataset



(a) Chilean dataset at HS 6-digit level



(b) Paraguayan dataset at HS 6-digit level



(c) USDA organic product dataset at HS 6-digit level

Source: Authors' calculations based on data from Chilean customs, Paraguayn customs, and USDA.