NLP ALGORITHMIC SOLUTION FOR GOODS CATEGORIZATION IN ACCORDANCE WITH THE COMMODITY NOMENCLATURE OF FOREIGN ECONOMIC ACTIVITY GIVEN TAX DOCUMENTS: A CASE STUDY IN ARMENIA

Zaruhi NAVASARDYAN
Ph.D. applicant, Mathematical Modeling of Economy, ASUE

Key words: taxation system, natural language processing, goods classification, keyword extraction, text similarity matching.

Introduction. Between and within country trade is subject to various taxes. Given that taxes apply based on the categories of good in trade, taxpayer have incentives to internationnally misclassify good for economic gains. Thus, the standardization of good classification is a key to correct tax collection. This consistency is pivotal in reducing ambiguity, minimizing errors, and fostering a shared understanding among stakeholders involved in trade. In the realm of international trade, customs data often stands as a paragon of meticulous categorization, adhering to the specific guidelines. The unified categorization approach is not merely a technical advancement; it is a transformative force that harmonizes the language of trade, ensuring a standardized understanding of goods across borders. This, in turn, facilitates smoother international transactions, reduces errors in reporting, and bolsters the overall integrity of trade data. However, a critical gap emerges within the internal trade dynamics, specifically in transactions between taxpayers using invoices and the final customer, documented through receipts. While these documents are carefully collected by the tax authorities and meticulously describe the products, they frequently lack the systematic categorization found in customs data.

Currently in Armenia, there is an issue of proper matching of customs data with tax returns data. Obviously, the problem exists not at the taxpayer aggregate data level, but at a more granular level. As tax legislation in the Republic of Armenia does not require unified classification for internal trade purposes. In particular, the customs data includes detailed coding of goods which are based on standard approach (in particular, the Eurasian Economic Union Commodity Nomenclature of External Economic Activity at 10-digit level). The application of product codes in tax documents (invoices, fiscal receipts) instead is optional (or fragmented). The only remaining option to track the goods within the supply chain is to use the textual unstructured data, which brings up a serious complication due to high variety of possible inputs for the same item. This research aims to apply a combination of machine learning and analytical tools to automate the product classification and matching to respective unified taxonomy. The economic purpose of the research is making the huge textual unstructured data, now subject only for limited manual analysis, easily usable for analytics and thus contribute to better fraud detection system. Different natural language processing (NLP) approaches were applied to classify the

textual data into product categories and items. Given the limited human resources, only a limited set of products undergo manual classification and analysis at SRC, the automation of the process by the application of NLP techniques will give the chance to extend the classification for a bigger set of good with no or very small adjustments.

The insights and methodologies and results presented in this paper lay the foundation for further analysis of the internal trade patterns and the make it possible to track the flow of the goods from import to final consumer. The practical contribution of the research is its direct applicability in planning and implementing targeted tax audits and upgrading the current methods applied by tax authorities. It will also contribute to the fulfilment of the requirements imposed by the Eurasian Economic Union towards increasing the tracking capacity of the customs and tax authorities. Obviously, the challenge of accurate goods classification is compounded by several factors:

Armenian language used. Modern natural language processing models especially transformer based large language models demand huge textual datasets and substantial computational resources for effective training. However, Armenian stands out as a low-resource language, compounding the difficulty of acquiring sufficient data. This unique linguistic context presents a distinctive challenge, as, at the time of this research, the unavailability of pretrained models capable of delivering high-quality results in Armenian makes the task of goods categorization more complicated and highlights the need for more innovative approaches.

The product description filled in manually. This manual input introduces a layer of subjectivity and variability, as taxpayers enjoy complete freedom in crafting these descripttions. Consequently, these descriptions may lack specific information about the product, leading to a potential loss of crucial details. Moreover, the absence of standardized abbreviations and the allowance for grammatical errors further contribute to the complexity of the categorization process.

The rest of this paper presents the following: methodological approach, highlighting both machine learning models applied and steps necessary to operationalize them in practice, data overview, and the results, followed by brief conclusions.

Methodology. The data for 2019 was extracted from SRC database ans used for analysis. The main resources used for research purposes:

- Import data with labeled product codes (11 digits), product definitions;
- Tax receipt data with human labeled 4-digit ADG codes (may not be 100% correct) and product definitions;
- Invoice data with product descriptions only (textual data).

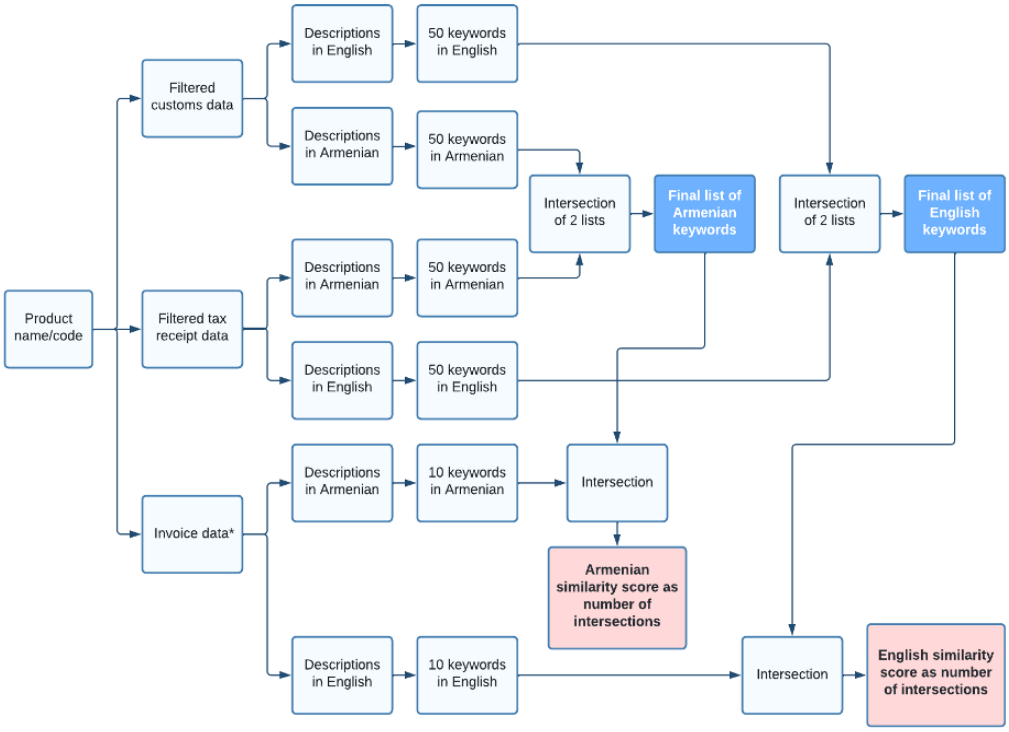
The data first underwent deep descriptive analysis, which includes checking the number of unique descriptions, number of unique importers of the chosen good, understandding the frequency and volume of imports, comparing unit prices. For some of these

goods the data is cross-checked with alternative sources to make sure no material errors are present. Considering that the data subject to analysis operates at the product level, it is essential to acknowledge that each individual invoice or tax receipt typically encompasses multiple products. This characteristic leads to a substantial increase of data points within the resultant dataset. Given the limited human and computation resources, the research was framed in the context of a small set of products, which would later serve as a base for an automated framework applicable to the full list of products. The list of products was chosen by compromise between complexity and verifiability (including using expert opinions). As all textual descriptions in the data are in Armenian which makes the analysis and the usage of state-of-the-art ML approaches complicated because of a very limited NLP research available in Armenian, thus all the textual information required for the task was translated into English using the Google Translation API. The main approach used for product categorization was text similarity approach based on the descriptions for imported goods. The approach consists of 2 main components: first is the keyword extraction and second the text similarity capturing. For keyword-based similarity identification unsupervised keyword extraction algorithm YAKE [Campos et al., 2018, 806–810] was used. This approach was chosen based on experimentation and result comparison between several approaches including TextRank a graph based method leveraging order in text for text summarization and keyword extraction [Mihalcea et al., 2004] and Rake. While Thushara et al. [2019, 969-973] get the best score for document keyword extraction performance with Rake, Yake is the faster and simpler alternative with similar performance. It is a frequency based automatic keyword extraction framework. The main advantage of the algorithm is that it does not use any fixed dictionary or 3rd source information for extraction, and it relies on features extracted from text, making it applicable to textual documents in different languages and domains without the need of prior domain knowledge. As the algorithm is language independent, it was applied to Armenian versions of product descriptions as well. First, keywords from text (both from Armenian and English) were extracted, then were later used instead of full descriptions.

The list representing the intersection of top 50 keywords extracted from labeled descriptions in customs and tax receipts data was built. Then for each product definition in invoice data a similarity score was calculated as the number of intersections or the number of keywords with similarity above threshold for Armenian and English respectively. To calculate similarity between 2 text entities the text is transformed into vectors and then cosine similarity between 2 vectors is computed. The second component of the similarity approach was the full description similarity estimation. For this purpose, Bidirectional Encoder Representations from Transformers - BERT [Devlin et al., 2019] model was used. Transformer architecture is one of the latest achievements of researchers in the deep learning sphere and is widely applied in natural language processing. A pre-trained BERT model developed and pre-trained by Google on a huge textual corpus and de-

signed in a way to take into account the context and to be able to learn contextualized embeddings was used.

Figure 1. Retrieving the text similarity from various transaction documents.



The other reason to choose this model for the task was the extensive use of BERT and its predecessor models for product description classification in the literature. Specifically, Jahanshahi et al. (2021) aimed to classify the product descriptions of several top online grocery platforms in Turkey. They compared the traditional methods such a bidirectional LSTM with more advanced NLP language models (BERT, ROBERTA etc.). While none of the techniques was an absolute winner, the advanced techniques outperformed in most cases. In ProBERT: Product Data Classification with Fine-tuning BERT Model [Zahera et al., 2020, 3] the authors fine-tuned the pre-trained BERT model for multiclass classify-cation and used it for the end-to-end classification. The advantage of the approach is that it solves the task with a single model and ProBERT will provide the class given the desc-ription, but it is not scalable and will need repetitive training whenever new ca-tegories appear. The similarity-based approach helped to exploit the ability of the pre-trained BERT to provide representative embeddings for product descriptions and bypass the sca-lability issue. As a result, 3 similarity scores are generated, among those are:

- Intersection of keywords in Armenian. For this point only the invoice descriptions ha-ving more than 1 keyword that match the keywords in the extracted list are assigned the given product class.

- Similarity of keywords in English. The criteria of similarity for this approach is to have at least one keyword from the keywords list that was built on labeled description translated into English.

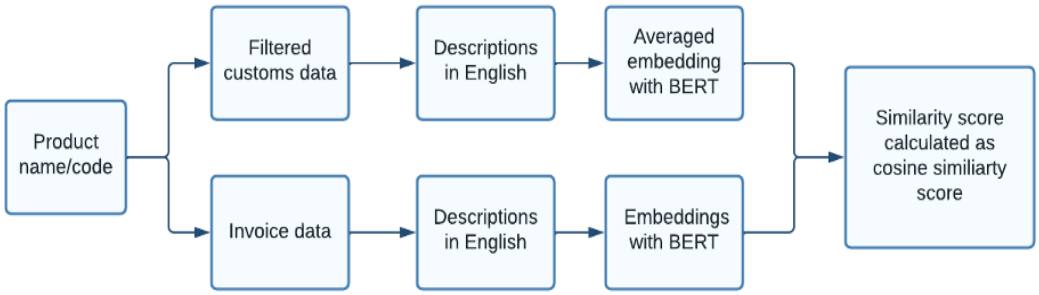
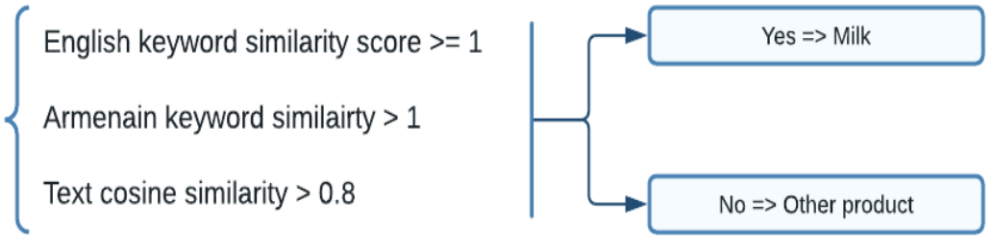


Figure 2. Using BERT to derive cosine similarity score.

- Similarity of full description. For similarity estimation of full descriptions, cosine similarity was used as a metric and the similarity threshold equal to 0.8 was taken as the optimal one for text matching[1].

Finally, the description was categorized as a specific type of product only if the above-mentioned 3 criteria are satisfied altogether. The decision criteria are summarized in the figure below.

Figure 3. Categorizing product as milk based on 3 criteria.



*Note: The similarity threshold can be adjusted based on some labeled examples.

Literature review. With the technological progress and given the level of automation that artificial intelligence can provide machine learning become more and more applicable in the state level systems. The applicability of such algorithms is especially high for tax authorities as the data is fully collected and documented in databases. While the usage of structured data is much simple, important information is often hidden in unstructured textual information. In their study on the data from United States Department of Agriculture researchers showcase the valuable role that artificial intelligence can play in

[1] The similarity score varies between 0 and 1. The threshold of 0.8 was identified based on manual check of the description after applying the threshold and considering that it is better to miss several milk products (false negatives) then to misclassify many other products as milk (false positives).

automating product classification at scale and, more generally, in empowering decision-makers with the information from unstructured data [Marra de Artiñano et al., 2023, 404-411]. In their study Chen et al. highlight the importance of correct product classification as inaccuracies not only pose duty risks but also undermine revenue collection efforts. They present an automated solution leveraging machine learning techniques to alleviate tand introduce a novel model based on off-the-shelf embedding encoders to cvalidate product classification in tax documents. On the human labeled data from Dutch customs officers they achieve 71% accuracy which is a promising result [Chen et al., 2021, 8-9]. Another approach used by researcher for text categorization is keyword extraction and matching. In their study named Domain classification of research papers using hybrid keyphrase extraction method researchers use RAKE: Rapid Automatic Keyword Extraction Algorithm for keyword extraction to classify the domain of documents and get interesting results [Thushara et al., 2018, 387–398].

Scientific novelty. The main scientific novelty of this applied research is the contribution to the field of good categorization by introducing and implementing Natural Language Processing (NLP) algorithmic solutions, particularly in the context of Armenia, where the challenges of a low-resource language and manual input of product descriptions on invoices and receipts have hindered the application of advanced NLP models.

Analyis. Withing the scope of the research, the experiment was meticulously executed, focusing on two key products: milk and rice. The selection of these products was guided by a comprehensive assessment of various characteristics inherent to them. This decision was carefully made through a thorough descriptive analysis and the incorporation of expert opinions. Multiple attributes were considered, such as the number of importers, the diversity within product descriptions, and the potential for false positives, considering the inherent nature of each product.

The evaluation of the NLP algorithmic solutions encompasses two pivotal components:

Human expert evaluation

A cadre of human experts, drawn from the State Revenue Committee (SRC) and independent researchers, meticulously conducted a manual evaluation of the classification results. The outcomes underscored the efficacy of the NLP algorithms, showcasing an average alignment of approximately 90% with human expert labeling. It is essential to note that labeling errors may occur in human labeled data as well. The results not only highlight the algorithm's prowess but also emphasize the importance of considering human opinion in the evaluation process.

Evaluation on test sets from human labeled tax receipt data and customs data.

To evaluate the algorithm's performance without direct human intervention, two distinct test sets were crafted, each representing 20% of taxpayer-labeled tax receipt and import data descriptions, respectively. The tax receipt labeling, marked by its lack of cont-

rol, introduces the possibility of labeling errors. On the other hand, the diversity within the textual content renders the evaluation more realistic, capturing the intricacies of varied expressions. On the contrary, the customs data descriptions exhibit precision, making the meaning explicit and facilitating straightforward classification. Acknowledging the advantages and drawbacks inherent in both datasets, the experiment was rigorously evaluated on both fronts. The table below presents the results of three distinct approaches for milk products, offering a detailed glimpse into the algorithm's performance across diverse datasets with varying characteristics. Recall was used as an evaluation metric. Recall shows the proportion of identified relevant products in the whole set of relevant products. In this context, false positives can be important as well as different product types to classify will be all pooled in one set. True positive rate was chosen as the main metric of evaluation as false positives are easier to identify and filter out based on other sources of information such as average price, quantity etc.

Table 1. Evaluation of 3 matching approaches for milk.

Approach	Recall	
	Tax receipts	Custom's data
Keyword extraction and simple intersection	0.83	1
Keyword extraction and similarity with Glove	0.83	1
Text similarity with Bert	0.93	1

The findings from the first and second approaches exhibit a remarkable level of concordance across nearly all cases. However, a difference emerges with the text similarity-based approach. This variance signifies that while the first and second methodologies align closely, the text similarity-based approach introduces a distinctive perspective. This difference, rather than being a discrepancy, underscores the complementary nature of these approaches, hinting at their potential synergies when employed in ensemble. The slight variations revealed by the text similarity-based approach could serve as valuable insights, contributing to a more comprehensive understanding of goods categorization. The prospect of ensembling these approaches emerge as a strategic consideration, harnessing their collective strengths to refine and elevate the accuracy of the overall classification process. This dual-layered evaluation approach not only ensures a comprehendsive understanding of the algorithm's performance but also establishes a robust framework for validating its accuracy and applicability across diverse datasets. The strategic

selection of milk and rice as the experimental focus adds depth to the evaluation, considering the specific characteristics of these products and their implications for goods categorization within the Armenian trade context.

Conclusion. Thus, this study presents a pioneering exploration into the realm of goods categorization, employing innovative Natural Language Processing (NLP) algorithmic solutions within the intricate context of the Armenian tax system. The results of the experiment, validated through human expert evaluation and meticulous assessments on diverse datasets, showcase the efficacy of the NLP approaches. The achieved accuracy, averaging around 90% alignment with human expert labeling, demonstrates the robustness of the NLP algorithms, despite the challenges posed by the low-resource nature of the Armenian language and the uncontrolled tax receipt labeling. The successful application of these methodologies not only enhances goods categorization precision but also lays the foundation for broader implications in the taxation system.

Moving forward, several avenues for future research and development emerge from this study. Firstly, the methodology can be extended to encompass a broader spectrum of products, beyond the initial focus on milk and rice. This extension will provide a comprehensive evaluation of the NLP algorithms' adaptability and effectiveness across diverse product categories, contributing to the generalizability of the proposed solutions. Moreover, the construction of a product flow network stands out as a promising next step. This involves mapping the flow of various products within the trade ecosystem, starting from the import and following the path up to the final consumer. Such a network will not only enhance our understanding of goods movement but also open avenues for optimizing supply chain management and trade policies. In essence, this study serves as a stepping stone for advancing the field of goods categorization, and the outlined future work provides a roadmap for continued exploration, refinement, and innovation in this dynamic and critical domain. The insights derived from this study not only contribute to academic discourse but also hold the promise of substantial real-world impact in automated framework for tax control and fraud detection practices.

References:

1. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). Yake! collection-independent automatic keyword extractor. Lecture Notes in Computer Science, pages 806–810. https://doi.org/10.1007/978-3-319-76941-7_80
2. Chen, H., van Rijnsoever, B., Molenhuis, M., van Dijk, D., Tan, Y., & Rukanova, B. (2021). The use of machine learning to identify the correctness of HS code for the Customs Import Declarations. 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pages 8-9. https://doi.org/10.1109/dsaa53316.2021.9564203
3. Devlin, J., Ming-Wei, C., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL). https://arxiv.org/pdf/1810.04805.pdf

4. Jahanshahi,H., Ozyegen, O., Cevik, M., Bulut,B., Yigit,D.,Gonen,F., Başar A. (2021). Text Classification for Predicting Multi-level Product Categories. https://arxiv.org/abs/2109.01084v1
5. Marra de Artiñano, I., Riottini Depetris, F., & Volpe Martincus, C. (2023). Automatic Product Classification in International Trade: Machine Learning and Large Language Models. https://doi.org/10.18235/0005012
6. Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing Order into Texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411.
7. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162
8. Thushara, M. G., Krishnapriya, M. S., & Nair, S. S. (2018). Domain classification of research papers using hybrid keyphrase extraction method. Advances in Intelligent Systems and Computing, pages 387–398. https://doi.org/10.1007/978-981-10-8636-6_40
9. Thushara, M. G., Mownika, T., & Mangamuru, R. (2019). A comparative study on different keyword extraction algorithms. 3rd International Conference on Computing Methodologies and Communication (ICCMC), pages 969-973. https://doi.org/10.1109/iccmc.2019.8819630
10. Zahera, H., Mohamed, Sh. (2020). ProBERT: Product Data Classification with Fine-tuning BERT Model. www.researchgate.net/publication/344901824_ProBERT_Product_Data

Zaruhi NAVASARDYAN

NLP algorithmic solution for goods categorization in accordance with the commodity nomenclature of foreign economic activity given tax documents: a case study in Armenia

Key words: taxation system, natural language processing, goods classification, keyword extraction, text similarity matching

Well organized and transparent tax systems are essential for the proper functioning of any country's economy. These systems are responsible for collecting and processing tax-related data from individuals and businesses and ensuring compliance with tax laws and regulations. This article introduces a pioneering approach to goods categorization within the context of internal trade through invoices and tax receipts in Armenia, leveraging cutting-edge Natural Language Processing (NLP) algorithms. Focused on addressing the challenges presented by the low-resource Armenian language and lack of control on the product description provided by taxpayers, the study meticulously evaluates the performance of the NLP algorithms through a case study involving products sample such as milk and rice. The ensemblling approach, by combining the nuanced contextual understanding of semantics with the structural accuracy of syntax, showcase enhanced adaptability and resilience in the face of linguistic diversity. Results indicate an average alignment of approximately 90% with human expert labeling, showcasing the robustness of the algorithms in achieving precise goods categorization. This study contributes valuable insights and empowers tax authorities with practical tools and algorithms with direct implycations for improving the efficiency and transparency of internal trade transactions.