INVENTORY CLASSIFICATION WITH AI: EVALUATING HOW LARGE LANGUAGE MODELS ENHANCE CATEGORIZATION USING UNPSC CODES

Anmolika Singh and Yuhang Diao

Data Scientist, USA

ABSTRACT

Effective item categorization plays a crucial role in transforming unstructured datasets into organized categories, simplifying inventory management for businesses. However, this process is often subjective and lacks consistency across industries and typically requires extensive manual effort for implementation. The United Nations Standard Products and Services Code (UNSPSC) offers a standardized framework for inventory cataloguing. This study examines the use of Large Language Models (LLMs) to automate the classification of inventory data into UNSPSC codes as the chosen taxonomy based on item descriptions. It evaluates the accuracy and efficiency of LLMs when processing datasets that are large and diverse; and when focusing on a specific segment judging the effect of providing context to the LLM. The results demonstrate that LLMs can significantly reduce the manual workload while maintaining high accuracy of upto 90% at UNSPSC segment level when LLM is provided with context. These findings present LLMs as a scalable and efficient solution for businesses seeking to automate inventory management, with the potential for further improvement through advanced model architectures and refined prompt engineering.

KEYWORDS

Item categorization, Inventory management, UNSPSC codes, Natural Language Processing (NLP), Large Language Models (LLMs), Automation, Data classification, Inventory standardization, UNSPSC Codes, Prompt Engineering, Artificial Intelligence (AI)

1. INTRODUCTION

Businesses consistently strive to manage operations and policies for their Stock Keeping Units (SKUs) in areas such as pricing, prioritization, e-commerce, accounting, and resource distribution. As inventory scales, so do the associated time requirements and costs [16]. Different types of items necessitate tailored strategies, simplifying processing and analysis. Consequently, there is a growing need to group products into categories.

Organizing items into product categories is a key to effective inventory management, improving the efficiency of supply chain operations. This classification facilitates inventory tracking, demand forecasting, and strategic decision-making. In the modern digital business environment, the importance of standardized information extends beyond automated processes to encompass all electronically supported operations [3]. Electronic product catalogues have thus become integral to inventory management, serving as centralized hubs for regularly updated and easily distributable product data. These catalogues ensure accuracy and consistency throughout the supply chain [3]. Structured and categorized data benefit manufacturers, distributors, wholesalers, and resellers by optimizing processes and enabling informed business decisions.

Adopting common coding standards such as UNSPSC is instrumental in transforming raw data into coherent, valuable information [4]. Standardized product classifications provided by UNSPSC enhance data usability for tasks like expenditure analysis, e-commerce, logistics, and supplier sourcing [4]. This systematic approach to categorization bolsters inventory management by allowing organizations to monitor inventory levels efficiently and identify potential costsaving opportunities [4].

In this paper, we study the possible automation of this item categorization using a Large Language model such as OpenAI's GPT-4 [12]. Given the complexity inherent in categorization processes, employing LLMs through prompt engineering offers a promising avenue to streamline and enhance the accuracy of this task. LLMs mark a significant advancement in artificial intelligence and natural language processing, showcasing remarkable capabilities in comprehending and generating desired outcomes based on input.

This paper serves as an extended version of the conference proceedings originally published in the International Journal on Cybernetics & Informatics (IJCI) under the title Leveraging Large Language Models for Optimized Item Categorization Using UNSPSC Taxonomy [20]. The extended version expands upon the experimentation and methodologies presented in the original publication, offering a more comprehensive analysis of leveraging Large Language Models (LLMs) for automating item categorization using the UNSPSC framework. It delves deeper into the technical evaluations through additional experimental results and broader implications of employing LLMs for efficient and accurate inventory classification, aiming to provide enhanced insights for researchers and practitioners in the field.

2. LITERATURE REVIEW

The field of item categorization has evolved significantly, leveraging advancements in machine learning and natural language processing (NLP) to address the challenges of managing large and diverse inventories. This section reviews key contributions to the literature, including traditional approaches to inventory classification, the structure and application of the UNSPSC taxonomy, the transformative potential of large language models (LLMs) for automated categorization, and the emerging importance of prompt engineering in refining model outputs. By synthesizing prior research, this review highlights both the progress and ongoing challenges in the domain of item categorization.

2.1. Item Inventory Classification

Various approaches have been explored to automate the item classification problem. Shankar and Lin [15] have worked on applying supervised learning with known product categories and multiclass features to increase accuracy in the classification of products. However, this method worked on more broad generalized product categories. The results of Gottipati and Vauhkonen [5] experiments indicated that while a Chi-square model of feature selection worked with a small feature size set, for large feature set size Naïve Bayes gave the most accuracy with plain frequency-based Unigram model for feature extraction and LDA fared at an average level. In [19] the authors suggest the Attention CNN (ACNN) model, for large-scale categorization of product titles into 35 top-level categories.

2.2. UNSPSC Codes

The UNSPSC taxonomy is constructed as a tree structure with four levels called Segment, Family, Class, and Commodity [17] Every level has a textual description as well as unique two

digits. From the top of the tree to the bottom, the category of each layer becomes one step more granular [8]. Combining all these codes gives the unique UNSPSC code for the product.

1. Commodity: 12345678
2. Class: 12345600
3. Family: 12340000
4. Segment: 12000000

There have also been studies that focused particularly on classification using the UNSPSC code standard. In [1] the authors employed machine learning algorithms to develop the classification model with SVM having the best accuracy. In another study, the researchers suggested using special clustering considering individual factors such as input or output of service operations to categorize services using UNSPSC Codes [9]. Karlsson and Karlstedt [8] in the thesis suggest that the optimal learning method for training is Support Vector Machines using inverse frequency class balancing, extracting features from the brand and title of products to give them an accurate UNSPSC taxonomy. However, the results of the study were inconclusive. Wolin [18] classifies products in UNSPSC taxonomy by creating vectors based on words in each product category of the training set and computing the cosine similarity measure between an input product feature vector and all candidate categories.

2.3. Large Language Models

Recent advancements in Large Language Models offer a promising avenue for automated item categorization to be carried out efficiently and accurately. LLMs have become a powerful tool to apply Natural Language Processing (NLP) tasks and can be adapted for specific use case scenarios. In [6] the authors compare GPT-3 and their own model WHAM in the task of classifying job postings by the option of remote work at least one day per week.

In a research closer to our objectives of product classification [10] the GPT-3.5 model showed a high accuracy rate in categorizing products using Harmonized System (HS) nomenclature into duty tariff lines whereas traditional machine learning models performed well only with their training dataset.

2.4. Prompt Engineering

Prompt engineering is a critical technique for effectively utilizing large language models (LLMs). By crafting precise and contextually rich prompts, researchers can guide LLMs to produce accurate and relevant outputs. This process is essential for reducing ambiguity and enhancing the consistency of the model's responses. Brown et al. [2] demonstrated the efficiency of few-shot learning with carefully constructed prompts, highlighting that even a few examples can substantially improve model performance. Similarly, Raffel et al. [13] emphasized the importance of prompt format and structure in their exploration of transfer learning with the T5 model, showing that prompt variations can lead to different levels of success in text-to-text tasks. In the context of item categorization, prompt engineering helps address challenges such as ambiguous item descriptions and the need for domain-specific knowledge. Schick and Schutze [14] explored the use of Cloze-style prompts, where the model fills in blanks within a sentence, for few-shot text classification. This technique demonstrated how prompts that mimic natural language questions elicit more accurate responses from LLMs. This approach is particularly relevant for UNSPSC categorization, where items must be matched to precise codes based on nuanced descriptions. By leveraging structured, contextual, and multi-turn prompts, researchers can enhance the accuracy of UNSPSC code assignment, thereby improving data standardization and procurement processes across various industries.

3. METHODOLOGY

This section outlines the approach used to explore the effectiveness of large language models (LLMs) for UNSPSC-based item categorization. The methodology involves leveraging a publicly available dataset, implementing structured prompts, and conducting experiments to evaluate the model's performance across various parameters.

3.1. Dataset

We use a publicly available dataset of Purchase Orders from the State Government of California [11] for our experiments. This dataset provides comprehensive details about various purchase orders issued during the specified period. The original dataset consists of 32 columns, capturing a wide range of information, including:

- Item Details: Item name, description, quantity, unit price, and total price.
- Supplier Details: Supplier code, name, qualifications, and zip code.
- Classification Codes and UNSPSC: The classification codes and normalized UNSPSC associated with the items, categorizing them based on commodity, class, family, and segment.

The dataset includes the following unique classifications:

- 7,658 Commodity Codes
- 1,821 Class Codes
- 388 Family Codes
- 57 Segment Codes

This dataset is particularly rich in information, making it an invaluable resource for studying item UNSPSC categorization. The inclusion of detailed item descriptions and their associated UNSPSC codes makes it well-suited for analysing the performance of large language models (LLMs) in item categorization tasks.

For **Experiment 1**, we take a random sample of 50,000 purchase order entries from the dataset, which provides a diverse set of products and categories for evaluating UNSPSC classification.

For **Experiment 2**, we focus on segment 440000, which pertains to "Office Equipment and Accessories and Supplies." This subset contains 27,731 products specifically within this segment. This refined dataset allows us to test the model's effectiveness in a focused category, providing a more granular analysis of how large language models (LLMs) perform in segment-specific UNSPSC categorization tasks.

3.2. Experiment Setup

3.2.1. Overview

The primary goal of this experiment is to leverage large language models (LLMs) such as Open AI's GPT-4 to classify items based on their names and descriptions using the UNSPSC. We evaluate the GPT-4 model's performance by comparing its classifications against a test dataset from the California State Government's purchase order data. We then test a subset of the data selecting one segment of items to compare the performance.

3.2.2. Dataset Preparation

The dataset is pre-processed to remove any incomplete records and to normalize item descriptions through methods such as text cleaning, and standardization.

3.2.3. Model Selection

The Azure Open AI API, specifically the GPT-4 model, is chosen for its advanced natural language processing capabilities. The GPT-4 model had the highest reasoning benchmarks at the time we ran the experiment [12]. It is accessed using an API key, with requests made to the specified endpoint.

3.2.4. Experimental Procedure

The experimental procedure involves a series of steps designed to leverage the Azure Open AI API for UNSPSC item categorization. Initially, input construction is performed for each item in the dataset, where a prompt is crafted to include both the item name and its description. These prompts are carefully formatted to ensure that the model receives all the necessary information for accurate classification. Following input construction, POST requests are sent to the API using these prompts. Each request is structured to query the model and retrieve a prediction for the UNSPSC code corresponding to the item. Once the responses are received, they are parsed to extract the predicted UNSPSC codes. This parsing process ensures that the outputs are correctly interpreted and stored for subsequent evaluation against the actual UNSPSC codes in the dataset. This systematic approach allows for a thorough assessment of the model's performance in classifying items based on their names and descriptions.

3.2.5. Prompt Building

For the study, we build a generic prompt that guides the LLM to assign UNSPSC Codes from the product name and item descriptions.

Prompt 1 - Generic Prompt
You will receive a product name and description. Your task is to classify the product into the appropriate UNSPSC category. Provide your output as the UNSPSC code only.

Next, we implemented a Cloze-style prompt, a widely used technique in natural language processing, where part of the input is left blank for the model to fill. This style is effective for classification tasks because it frames the problem as a question with a single, context-dependent answer. In our case, we structured the prompt to mimic a natural query for the UNSPSC code, expecting it to improve accuracy by reducing ambiguity in the model's output.

Prompt 2 - Cloze-style Prompt
The appropriate UNSPSC code (a numerical code) for a product named ‘{item_name}’ described as ‘{item_description}’ is:"

In another approach, we used a long prompt-engineered statement that provides context to the LLM by including example usage. The following are examples of product names and descriptions included in the user prompt to guide the model.

Prompt 3 - Engineered Prompt
You will receive a product name and description. Your task is to classify the product into the appropriate UNSPSC category. Provide your output as the UNSPSC code only.

Example 1:
User Prompt:
Product Name: HP LaserJet Pro M404dn
Description: Laser printer, black and white, 40 pages per minute.
Expected Output: 43212110

Example 2:
User Prompt:
Product Name: Dell Latitude 7420
Description: Business laptop with 14-inch screen and Intel i7 processor.
Expected Output: 43211503

Example 3:
User Prompt:
Product Name: 3M Scotch Magic Tape
Description: Invisible tape for office use, 1 inch by 1000-inch roll.
Expected Output: 31201512

With these three prompts, we run our experiments to test the impact of prompt creation on the accuracy of the LLM.

For the second part of our experiment, all three prompts are tweaked to provide relevant context to the LLM with the dataset limited to one segment of 440000, which pertains to "Office Equipment and Accessories and Supplies."

For the generic Prompt, the LLM is guided to classify ‘office supplies’ that provide more context.

Prompt 1.1 - Generic Prompt:
You are an expert in product categorization with a specialization in office supplies. You will be provided with a product name and description. Your task is to determine the most accurate UNSPSC code specific to office supplies for the given product. Output only the UNSPSC code, without any additional text or explanation.

Again, for the Cloze-Style prompt, the LLM is told that the item is an office supply product.

Prompt 2.1 - Cloze Prompt
For an office supply product named '{item_name}', described as '{item_description}', determine the most appropriate UNSPSC code (a numerical classification code):

In the last long prompt-engineered statement we provided context to the LLM by telling it to be an expert in office supplies and including examples specific to the segment. The following are examples of product names and descriptions included in the user prompt to guide the model.

Prompt 3.1 - Engineered Prompt
You are an expert in product categorization with a specialization in office supplies. You will be provided with a product name and description. Your task is to determine the most accurate UNSPSC code specific to office supplies for the given product. Output only the UNSPSC code, without any additional text or explanation.

Example 1:
User Prompt:
Product Name: ACCO Jumbo Paper Clips,
Description: 'Non-skid jumbo paper clips, 100 count.
Expected: 44122104

Example 2:
User Prompt:
Product Name: Swingline Desktop Stapler
Description: Standard-size stapler for 20 sheets with metal construction
Expected: 44121615

Example 3:
User Prompt:
Product Name: Expo Dry Erase Markers
Description: Assorted color dry erase markers for whiteboards, chisel tip, pack of 8.
Expected: 44121708

3.3. Result Metrics

In each experiment, we test the model's performance using three types of prompts: a generic prompt, a Cloze-style prompt, and an engineered prompt. For each prompt type, the model is evaluated at three distinct temperature levels. The temperature parameter in LLMs regulates the variability of the output. Higher temperature values encourage the model to explore a wider range of possibilities, resulting in more creative and diverse responses. However, this added diversity may reduce accuracy or lead to inconsistent predictions. On the other hand, lower temperature values prioritize precision and consistency by minimizing randomness. Since the goal of this experiment is to perform a classification task, it is expected that lower temperature settings will deliver higher accuracy, as the outputs are more deterministic and reliable.

Performance evaluation is based on accuracy, defined as the proportion of correct product categorizations made by the model. Accuracy is calculated for each hierarchical level of the UNSPSC code: Commodity, Class, Family, and Segment. Furthermore, in the second experiment, we extend the evaluation to a smaller subset of the original dataset, specifically focusing on a single segment of UNSPSC codes: 440000, which pertains to "Office Equipment and Accessories and Supplies." This additional analysis provides a more detailed understanding of the model's performance within a specific product category. The experiment measures performance across three key factors:

1. Type of Prompt
2. Temperature Level
3. UNSPSC Hierarchy Level

4. RESULT ANALYSIS

The experiment evaluated the prediction accuracy of LLMs, to classify items into UNSPSC codes at different levels of the hierarchy, ranging from commodity to segment level. The results, as presented in Table 1, Table 2, and Table 3, provide insights into the performance of different prompts and temperature settings on the full diverse dataset, and Table 4, Table 5, and Table 6, provide insights into the performance of different prompts and temperature settings on the subset dataset belonging to one UNSPSC segment.

Across all tested prompts, it was observed that the prediction accuracy generally increased as we moved up the hierarchy of UNSPSC code from commodity to segment level. Also, a temperature of '0' consistently yielded the best accuracy across all the results. This suggests that a lower temperature parameter results in more deterministic and accurate predictions.

Amongst the 3 tested prompts, Prompt 3 ‘Engineered Prompt’ consistently performed the best in terms of accuracy for the full dataset. Prompt 3 was particularly effective in handling cases where the model encountered inadequate information, as it provided a statement informing about the insufficiency rather than supplying a potentially incorrect result. Whereas for the second part of the experiment, where the dataset is limited to a single segment (440000: "Office Equipment and Accessories and Supplies"), significantly reducing variability in the data, Prompt 1 ‘Generic Prompt’ gave the highest accuracy.

Table 1. Prompt 1 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	9.73%	24.88%	34.72%	49.09%
0.50	10.09%	25.64%	36.08 %	49.59%
0	10.20%	25.88%	35.75%	49.88%

Table 2. Prompt 2 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	11.07%	27.44%	38.80%	53.20%
0.50	11.29%	28.01%	39.27 %	53.63%
0	11.45%	28.30%	39.52%	53.73%

Table 3. Prompt 3 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	10.34%	27.74%	39.00%	53.10%
0.50	10.73%	28.77%	40.23 %	54.49%
0	10.80%	29.01%	40.31%	54.59%

Table 4. Prompt 1.1 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	47.14%	71.95%	82.08%	89.87%
0.50	47.61%	72.30%	82.41%	90.14%
0	48.05%	72.60%	82.66%	90.33%

Table 5. Prompt 2.1 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	45.00%	69.96%	79.85%	87.54%
0.50	45.85%	70.70%	80.41%	87.98%
0	45.98%	71.05%	80.64%	88.22%

Table 6. Prompt 3.1 Matrix

Temperature	Accuracy Commodity	Accuracy Class	Accuracy Family	Accuracy Segment
1	41.14%	69.25%	78.40%	85.91%
0.50	41.63%	69.73%	78.75%	86.19%
0	41.75%	69.79%	78.67%	86.12%

Overall, the experiment results highlight the importance of prompt design and temperature setting in achieving accurate UNSPSC code predictions. The findings suggest that utilizing Prompt 3 with a temperature of '0' can lead to improved accuracy, particularly at higher levels of the UNSPSC hierarchy for a dataset that is diverse and when the context is unknown and when the dataset belongs to a niche with context known Prompt 1 with a temperature of '0' can lead to improved accuracy These insights can be valuable for researchers and practitioners working on enhancing the prediction accuracy of UNSPSC codes.

5. CONCLUSION

In this research, we conducted an extensive experiment to evaluate the efficiency of LLM's capability to categorize items into UNSPSC codes. Our findings reveal that prediction accuracy generally improves as the task transitions from the commodity to the segment level, highlighting the role of granularity in enhancing model performance. Additionally, a temperature setting of ‘0’ consistently outperformed higher temperature settings, underscoring the importance of determinism for tasks requiring precision. Interestingly, while an engineered prompt demonstrated superior performance in the full dataset due to its ability to handle uncertainty effectively, a generic prompt achieved the highest accuracy within the subset data. This divergence can be attributed to the reduced variability and complexity in the subset, which favoured a simpler and more direct approach. Overall, our experiment with the subset dataset, which focused on a single segment, yielded high performance. This finding suggests that providing LLMs with more explicit, targeted information about the dataset can lead to even more accurate results.

When comparing our results to earlier work on the classification of products and services using UNSPSC codes [7], several key observations emerge. The baseline method utilized Naive Bayes

probabilistic models and achieved promising results at the class level. Subsequent improvements using word2vec with averaging and logistic regression, as well as FastText and RoBERTa, demonstrated enhanced performance, with RoBERTa achieving the highest mF1 and wF1 scores of 0.912 and 0.904, respectively. This dataset consisted of 30,000 samples spread across 46 UNSPSC categories. However, the authors reported a significant drop in performance to mF1 scores of 50%-55% when tested on a larger, more diverse dataset.

In contrast, our experiment utilized a larger dataset of 50,000 data points, which was more diverse and spanned a broader range of UNSPSC categories. Within this dataset, the accuracy at the class level reached 40.31%, while the segment level achieved 54.59%, demonstrating better performance at higher levels of the UNSPSC hierarchy. For the smaller subset of one segment (27,000 data points), accuracy significantly improved to 72.60% at the class level and 90.33% at the segment level.

In conclusion, LLMs are on power with previously tested classification models. Hence, LLMs have proven to be a highly effective resource for item categorization, significantly reducing time and manual effort. Our research contributes valuable insights into the effectiveness of LLMs emphasizing the significance of prompt design and temperature settings.

In general, the success of LLMs in this domain indicates their potential as powerful tools for item categorization across various taxonomies. As these models become more sophisticated, they could revolutionize the way organizations approach classification tasks, offering efficient and scalable solutions for managing large datasets and complex taxonomies. This could lead to broader applications in industries that rely heavily on accurate and efficient categorization, ultimately enhancing productivity and decision-making processes.

6. FURTHER WORK

The rapid evolution of Large Language Models (LLMs) opens new avenues for advancing item categorization tasks. Our experiment utilized Open AI's GPT-4, recognized as the best performing model available at the time of this study, ensuring our findings were benchmarked against the most capable technology of the period. However, with the continuous release of more advanced models and other emerging architectures, it is crucial to revisit these experiments with newer LLMs. These upcoming models are expected to offer enhanced contextual understanding, improved reasoning capabilities, and greater adaptability to diverse datasets, potentially surpassing the results achieved in this study.

While this research focused on UNSPSC codes, its methodologies are transferable to other classification systems. Testing newer LLMs on frameworks like the North American Industry Classification System (NAICS) or Harmonized System (HS) codes could demonstrate their adaptability and broader application. Moreover, hybrid approaches that integrate traditional machine learning models with LLMs could leverage the strengths of both paradigms to further improve classification efficiency and accuracy.

Our experiment’s finding suggests that providing LLMs with more explicit, targeted information about the dataset can lead to even more accurate results. It would be valuable to examine if this approach holds across other domains and datasets, and to determine if segment-based performance remains a key factor in overall accuracy.

ACKNOWLEDGEMENTS

We would like to express our gratitude to everyone who provided support and guidance throughout the development of this research.

REFERENCES

[1] Bello Abdullahi, YahayaMakarfi Ibrahim, Ahmed Doko Ibrahim, KabirBala, Yusuf Ibrahim, and Muhammad AliyuYamusa. Development of machine learning models for categorisation of nigerian government's procurement spending to unspsc procurement taxonomy. International Journal of Procurement Management, 19(1):106–121, 2024.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, PrafullaDhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[3] Frank Brüggemann and Ursula Hübner. From Product Identification to Catalog Standards, pages 127–154. Springer London, London, 2008.

[4] A.M. Fairchild and B. de Vuyst. Coding standards benefiting product and service information in e-commerce. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, pages 3201–3208, 2002.

[5] Srinivasu Gottipati. E-commerce product categorization srinivasugottipati and mumtazvauhkonen, 2012.

[6] Stephen Hansen, Peter John Lambert, Nicholas Bloom, Steven J Davis, RaffaellaSadun, and BlediTaska. Remote work across jobs, companies, and space. Technical report, National Bureau of Economic Research, 2023.

[7] Ihor Hrysha and Samuel Grondahl. Large-scale product classification to recommend suppliers in procurement systems.

[8] Mikael Karlsson and Anton Karlstedt. Product classification-a hierarchical approach. LU-CS-EX 2016-31, 2016.

[9] Qianhui Liang, Peipei Li, Patrick CK Hung, and Xindong Wu. Clustering web services for automatic categorization. In 2009 IEEE International Conference on Services Computing, pages 380–387. IEEE, 2009.

[10] Ignacio Marra de Artiñano, Franco RiottiniDepetris, and Christian Volpe Martincus. Automatic product classification in international trade: Machine learning and large language models. Technical report, IDB Working Paper Series, 2023.

[11] State of California. Purchase order data, 2024.

[12] OpenAI. gpt-4-0613. https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4, 2023. Accessed: June 11, 2024.

[13] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, SharanNarang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67, 2019.

[14] Timo Schick and HinrichSchütze. Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676, 2020.

[15] SushantShankar and Irving Lin. Applying machine learning to product categorization. Department of Computer Science, Stanford University,2011.

[16] BanuSoylu and BaharAkyol. Multi-criteria inventory classification with reference items. Computers & Industrial Engineering, 69:12–20, 2014.

[17] unspsc org. United nations standard products and services code® (un-spsc®), 2022.

[18] Ben Wolin. Automatic classification in product catalogs. In Proceedings of the 25th annual international acmsigir conference on research and development in information retrieval, pages 351–352, 2002.

[19] Yandi Xia, Aaron Levine, Pradipto Das, Giuseppe Di Fabbrizio, KeijiShinzato, and AnkurDatta. Large-scale categorization of japanese product titles using neural attention models. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 663–668, 2017.

[20] Singh, Anmolika & Diao, Yuhang. (2024). Leveraging Large Language Models For Optimized Item Categorization using UNSPSC Taxonomy. International Journal on Cybernetics & Informatics. 13. 10.5121/ijci.2024.130601.

AUTHORS

Anmolika Singh received the B.S. degree in Applied Data Sciences from the Pennsylvania State University, University Park, in 2021. From 2021 to 2024, she was part of the prestigious Stanley Leadership Program, where she served as an Artificial Intelligence and Data Analytics Associate working on multiple high impact projects. She currently serves as a Data Scientist at Stanley Black & Decker's Industrial Business Unit, where she extracts valuable insights from industrial data to inform strategic decision-making and process improvements. She is the author of several research articles, and her interests include applying data science techniques to solve real-world problems.

YuhangDiao earned a B.S. degree in Econometrics from Ohio State University in 2019, followed by a master's degree in Data Analytics Engineering from Northeastern University in 2021. From 2022 to 2024, he participated in the Stanley Black & Decker Leadership Program as a Data Analyst. During this time, he worked on developing predictive models and optimizing data-driven decision-making processes. Currently, he is a Data Scientist in the Industrial Segment, where he has implemented advanced machine learning techniques, such as classification algorithms and neural networks, to drive innovation and operational efficiency. His work has contributed significantly to the development of smart storage solutions, including automated inventory management systems that leverage AI to enhance accuracy and performance.