Detection of Vehicle Purchases in Various Invoices using Large-Scale Language Models

Gabriel V. Heisler[1], **William J. Beckhauser**[1], **Vitória S. Santos**[1], **Renato Fileto**[1]

[1]Department of Informatics and Statistics – Federal University of Santa Catarina,
Florianópolis, SC, Brazil

Corresponding e-mail: gvheisler@gmail.com

Abstract. *The precise identification of products in invoice item descriptions is crucial for applications such as auditing and fraud detection. However, the free-text descriptions of these items are often short, diverse, and inconsistent with other data fields. It makes product identification challenging and compromises the performance of existing solutions. In this work we compare the performance and computational costs of language models (LLMs) in the task of detecting vehicle descriptions in an invoice dataset from public purchases. Experimental results reveal that some state-of-the-art LLMs can reach high performance, even in noisy scenarios, and lightweight models yield competitive performance with lower computational costs.*

1. Introduction

The analysis of electronic invoice (NF-e) item descriptions to identify which products they refer to is a necessary step for audits, fraud detection, and promoting transparency in government spending [Hamdi et al. 2021, Saout et al. 2024]. Although the NF-e provides the GTIN *(Global Trade Item Number)* and NCM *(Mercosur Common Nomenclature)* fields, in public procurement datasets the GTIN is often missing or incorrect, while the NCM is usually available for each purchased item, since it is a mandatory field in the Brazilian NF-e data standard [Di Oliveira et al. 2024]. Thus, the NCM constitutes a pillar for invoice item classification – even though it refers to coarse categories of products and may also be incorrect or inconsistent sometimes.

The NCM is a code of up to eight digits that refer (from left to right) to increasingly refined classes. For example, the four-digit prefix 8703 designates "passenger automobiles and other motor vehicles mainly designed for the transport of people"; longer codes refine characteristics, such as 8703.24 ("engine displacement over 3,000 cm^3") and 8703.24.10 ("capacity of up to six people, including the driver"). In practice, inappropriate uses of NCM prefixes are observed; for example, 8703 is also frequently found in NF-e items that are not vehicles, but, in fact, vehicle parts, insurance, or scrap.

The state-of-the-art methods for automatic classification of invoice items based mainly on their free-text product descriptions combine distributive (vector) representations of product descriptions, such as TF-IDF and *embeddings* [Gasparetto et al. 2022, Da Costa et al. 2023], with clustering algorithms or classical machine learning (ML) models, such as SVM, Random Forest, and Neural Networks. Recently, advances in Large Language Models (LLMs) have created new possibilities for attribute extraction and text classification tasks. However, there are challenges in scenarios with short, non-standardized, and noisy text, such as invoice item descriptions.

In this work we investigate approaches for classifying invoice items by product category. Our experiments compare the performance and computational costs of state-of-the-art language models (LLMs) in the task of detecting vehicle descriptions in a public procurement invoice dataset. The results show that Qwen 2.5, GPT-oss, Gemma 3, and Mistral maintain high accuracy even under noise, with Qwen 2.5 32B standing out (97.85%). In addition, more compact models, such as Qwen 2.5 7B, Gemma 3 4B, and Mistral 7B, achieve competitive performance with lower computational cost.

2. Related Work

Information extraction from text documents like invoice item descriptions, and their classification by product (category), are crucial for downstream applications like auditing and overpricing detection [Silva et al. 2023, Soares et al. 2024]. These problems have been widely studied [Sarawagi et al. 2008, Holt and Chisholm 2018].

However, challenges such as inconsistencies, typos, abbreviations, and incomplete descriptions [Saout et al. 2024, Hamdi et al. 2021] make the direct use of traditional Natural Language Processing (NLP) techniques difficult. On the other hand, methods for classifying NF-e items that combine vector representations with supervised classifiers such as SVM, Random Forest, and neural networks [Bardelli et al. 2020, Tutica et al. 2020] are strongly dependent on data quality, which is not always guaranteed in public procurement. Thus, more robust and versatile are needed.

The advancement of LLMs has opened up new opportunities for recognizing and standardizing product attributes in free text [Brinkmann et al. 2024]. However, the challenge of identifying products in short, noisy, and ambiguous texts, common in invoices, requires research comparing alternative approaches, including various LLMs.

3. Experiments

The proposed process, shown in Figure 1, consists of the following steps: (i) filtering items by the general NCM category (e.g. 4-digit prefix denoting passenger motor vehicle); (ii) using LLMs to classify items as actually belonging or not to a desired category; and (iii) evaluating the obtained results.

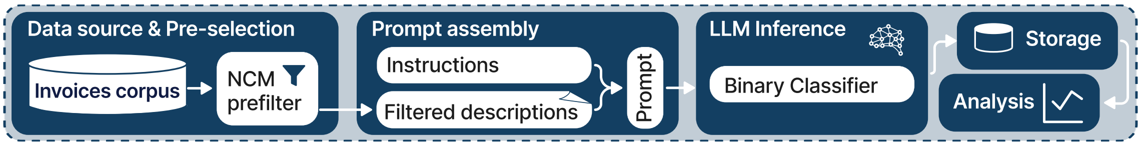

Figure 1. Proposal to determine if invoice items belong to a product category.

The experiments were conducted on a dataset of electronic invoices from public procurement. A filter by NCM was applied, totaling more than 42,000 item descriptions. For this study, we performed a frequency-weighted random sampling of product descriptions following the methodology from [Rea and Parker 2012], ensuring statistical validity. With a 99% confidence level and a 5% margin of error, the resulting sample contains 654 item descriptions manually labeled as *vehicle* or *non-vehicle*. The class distribution in the annotated set was 66.2% *vehicles* and 33.79% *non-vehicles*. Examples of descriptions:

- Vehicle: "vw gol 1.0 flex 4 portas"
- Non-vehicle: "motor de carro", "ford fiesta aluguel", "sucata gol 1.0 flex 4 portas"

4. Results

Table 1 summarizes the LLMs' accuracy, average latency, average number of tokens (input + output) per item processed, and throughput (number of tokens per second). Note that *qwen2.5:32b*, *gpt-oss:20b*, and *gemma3:27b* achieve accuracies close to or greater than 96%. Lighter models – especially *qwen2.5:7b* – also perform well, but with significantly shorter inference times and token consumption similar to other models. For comparison, we also conducted experiments with classical ML algorithms using TF-IDF features, with default `scikit-learn` parameters and a 70/30 train/test split. Random Forest (94.9% accuracy), SVM (93.9%), Naive Bayes (93.9%), and Logistic Regression (91.8%) were the best performing ones. On the other hand, surprisingly, state-of-art models like qwen3 and deepseek-r1 performed poorly. Due to space limitations we discuss these issues and examples of wrong results at GitHub[1].

Table 1. Accuracy in the identification of items from NF-es referring to vehicles

Model	Accuracy	Avg. Latency (s)	Avg. Tokens	Throughput
qwen2.5:32b	**97.85**	0.99	155.10	155.26
qwen2.5:7b	96.94	0.32	155.10	480.18
gpt-oss:20b	96.33	30.52	688.50	22.55
gemma3:27b	96.17	1.10	120.90	109.41
gemma3:12b	91.74	0.69	120.90	173.96
gemma3:4b	89.45	0.43	120.90	279.86
mistral:7b	88.22	0.65	163.20	249.54
qwen2.5:1.5b	80.88	0.21	155.10	731.60
qwen2.5:3b	76.91	0.27	155.10	568.13
qwen2.5:0.5b	66.66	**0.17**	155.10	**896.53**
llama3.2:3b	49.84	0.27	150.00	543.48
gemma3:1b	38.22	0.30	120.90	401.66
qwen3:32b	33.79	76.83	478.50	6.22
deepseek-r1:8b	33.79	18.32	470.10	25.65
qwen3:8b	33.79	16.50	466.30	28.26

5. Conclusions and Future Work

This work presented a study on the classification of invoice item descriptions for vehicle identification. The results indicate that LLMs can achieve high accuracy even in noisy scenarios, although lighter alternatives can also deliver competitive performance.

Future work includes experiments for identifying other product categories, multi-class classification, and expanding the proposed pipeline for hierarchical classification, in which identifying general product classes will be the first step before extracting attributes

[1] Code, prompt, and more technical details available at github.com/gvheisler/VehicleDetection

(e.g., brand, model, year) and classifying in more detailed categories or specific products. In addition, we plan to evaluate how NCM and GTIN codes can complement the classification process and assist in ground truth production, model training and evaluation.

Acknowledgments

This study was financed by the CAPES – Finance Code 001 – and by MPSC – Project Ceos: Data Intelligence for the Society. It also fits in the objectives of the INCT-tildIAr, financed by CNPq and FAPEMIG.

References

[Bardelli et al. 2020] Bardelli, C., Rondinelli, A., Vecchio, R., and Figini, S. (2020). Automatic electronic invoice classification using machine learning models. *Machine Learning and Knowledge Extraction*, 2(4):617–629.

[Brinkmann et al. 2024] Brinkmann, A., Baumann, N., and Bizer, C. (2024). Using llms for the extraction and normalization of product attribute values. In *ADBIS*, pages 217–230. Springer.

[Da Costa et al. 2023] Da Costa, L. S., Oliveira, I. L., and Fileto, R. (2023). Text classification using embeddings: a survey. *Knowledge and Information Systems*, 65(7):2761–2803.

[Di Oliveira et al. 2024] Di Oliveira, V., Bezerra, Y. F., Weigang, L., Brom, P. C., and Celestino, V. R. R. (2024). Slim-raft: A novel fine-tuning approach to improve cross-linguistic performance for mercosur common nomenclature. *arXiv preprint arXiv:2408.03936*.

[Gasparetto et al. 2022] Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.

[Hamdi et al. 2021] Hamdi, A., Carel, E., Joseph, A., Coustaty, M., and Doucet, A. (2021). Information extraction from invoices. In *ICDAR*, pages 699–714. Springer.

[Holt and Chisholm 2018] Holt, X. and Chisholm, A. (2018). Extracting structured data from invoices. In *Proceedings of the ALTA Workshop 2018*, pages 53–59.

[Rea and Parker 2012] Rea, L. and Parker, R. (2012). *Designing and Conducting Survey Research: A Comprehensive Guide*. Wiley.

[Saout et al. 2024] Saout, T., Lardeux, F., and Saubion, F. (2024). An overview of data extraction from invoices. *IEEE Access*, 12:19872–19886.

[Sarawagi et al. 2008] Sarawagi, S. et al. (2008). Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

[Silva et al. 2023] Silva, M. O., Costa, L. L., Bezerra, G., Gomide, L. D., Hott, H. R., Oliveira, G. P., Brandao, M. A., Lacerda, A., and Pappa, G. (2023). Análise de sobrepreço em itens de licitaçoes públicas. In *WCGE*, pages 118–129. SBC.

[Soares et al. 2024] Soares, D., da Silva, J. P. D., Zibetti, A. W., and Werner, S. S. (2024). Sobrepreço em compras públicas: Metodologia baseada na identificação de valores discrepantes. In *SBBD*, pages 266–272. SBC.

[Tutica et al. 2020] Tutica, L., Vineel, K., Mishra, S., Mishra, M. K., and Suman, S. (2020). Invoice deduction classification using lgbm prediction model. In *ETAEERE*, pages 127–137. Springer.