EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

Abstract
EMOS Master thesis competition 2025

'New Challenges in Official Statistics: Big Data Analytics and Multi-level Product Classification of Web-Scraped Data'

Author: **Juliana Machado, University of Porto**

Keywords: Big Data, Machine Learning, Official Statistics

1. INTRODUCTION

The rapid growth of data generated on the web offers unprecedented opportunities to modernize data collection methods across various domains within official statistics. As more detailed and timely information becomes available through these non-traditional data sources, there is significant potential to enhance traditional statistical indicators by integrating high-frequency and granular data. However, much of this web-generated data is unstructured, presenting unique challenges for official statistics. Without additional processing, it is not easy to utilize such data for precise statistical analysis and reporting. Banco de Portugal had been collecting retail data daily through web scraping from several retail brands, but this data was initially not usable for economic research because the products lacked official classifications. This exponential growth of collected data underscored the need to understand the state-of-the-art in Big Data technologies and, mainly, to automatize the classification of the collected products according to the European Classification of Individual Consumption According to Purpose (ECOICOP), thereby enabling these databases to be used for economic research. Therefore, the primary objectives of this study are to conduct a comprehensive literature review on Big Data technologies, explore the use of web scraping in official statistics, and examine techniques for multi-level product classification. Additionally, the study aims to develop a machine learning-based classification pipeline to automatically categorize web-scraped retail data into 71 categories of ECOICOP related to food and beverage. In addition to developing a practical classification pipeline, this study addresses technical questions essential for refining model performance and applicability. Specifically, it investigates which machine learning models are most effective for short-text classification at this specific case, and whether language-specific large language models outperform cross-language models in this context.

2. OBJECTIVES

The main objectives of this study are as follows:

1. **Conduct a Comprehensive Literature Review:** Explore state-of-the-art technologies related to Big Data, investigate the use of web scraping in official statistics, and examine the latest techniques for short text classification to be used in this research.
2. **Utilize Iterative Labeling and Model Training Processes:** Given the absence of a labeled dataset, develop an iterative approach that begins with manually labeling a small dataset to train initial models. These models are then used to label larger datasets iteratively, improving

the classification model's accuracy and efficiency over time.

3. **Develop a Machine Learning-Based Classification Pipeline:** Build a complete pipeline that applies Machine Learning techniques to automatically classify webscraped retail food and beverage data according to the European Classification of Individual Consumption According to Purpose (ECOICOP). This allows the web scraped data to be used for economic research and CPI nowcasting.
4. **Determine the Most Effective Model for Short Text Classification:** Evaluate various machine learning models to identify which is best suited for the classification of food and beverage products using only the name and brand.
5. **Compare Language-Specific and Cross-Language Models:** Investigate whether large language models specifically trained for Portuguese, such as BERTimbau, demonstrate superior performance compared to cross-language models like XLM-RoBERTa for this classification task

3. METHODS

This research employs a two-phase methodology: a literature review focused on Big Data technologies and techniques for multi-level product classification, followed by the implementation of machine learning models for product classification

Phase 1: Literature Review

The initial phase involved conducting a comprehensive literature review to understand the current state of Big Data technologies. Subsequently, research was undertaken to examine the use of web scraping in official statistics, providing a deeper insight into this topic. Finally, an exploration of the most suitable techniques for multi-level product classification was conducted:

1. **Big Data Technologies:** The literature review explored frameworks and methodologies relevant to managing large-scale data, including data storage and processing techniques. The aim was to assess how these technologies could be utilized to handle the exponential growth of data volumes at Banco de Portugal.
2. **Web Scraping in Official Statistics:** This section of the review focused on the role of web scraping as a non-traditional data collection method within official statistics.
3. **Multi-level Product Classification Techniques:** The literature review examined various techniques for classifying products at multiple levels. This classification presents distinct challenges in natural language processing, particularly due to the characteristics of short-text classification, which often suffers from a lack of context.

Phase 2: Machine Learning Implementation

The second phase focused on developing and evaluating machine learning models to classify web-scraped food and beverage retail data according to 71 categories of European Classification of Individual Consumption According to Purpose (ECOICOP). A significant challenge encountered in this phase was the absence of a pre-labeled dataset, which necessitated a careful approach to building an effective classification system. The following key steps were implemented:

1. **Addressing the Lack of Labeled Data:** To overcome the challenge of unlabeled data, an iterative labeling process was established. A small dataset was obtained from a single day's worth of web-scraped data from one supermarket, which, although limited in size, provided a diverse sample of food and beverage products. This dataset included various product categories and brands, creating a representative foundation for manual labeling. The labeled subset served as the starting point for training initial machine learning models.
2. **Data Preprocessing:** The dataset, provided by Banco de Portugal, consisted of food and beverage product titles, brands, and other relevant details scraped from various Portuguese supermarkets. Extensive preprocessing was conducted to clean and standardize the data, preparing it for classification. This involved removing noise, normalizing text, and tokenizing

product descriptions.

3. **Model Selection and Training:** Various machine learning models were evaluated for their effectiveness in short-text classification. Traditional algorithms such as Support Vector Machines (SVM) and XGBoost were tested alongside advanced deep learning models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and large language models, including BERT, XLM-RoBERTa, and BERTimbau. The latter, specifically designed for the Portuguese language, leverages transfer learning to enhance performance in classification tasks involving Portuguese text. Additionally, pre-trained word embeddings specific to Portuguese were incorporated to ascertain whether this strategy could further improve the effectiveness of the deep learning models. The goal was to determine the most effective approach for accurately classifying the short product titles in the food and beverage category.
4. **Model Evaluation:** To determine the most effective model, the dataset was strategically partitioned, with 80% allocated for model training and the remaining 20% designated for testing. The evaluation metrics for model performance included accuracy and the F1 Macro score. Accuracy measures the proportion of correctly classified 5 instances among the total instances, while the F1 Macro score evaluates the balance between precision and recall across all categories. This distinction is especially im- portant in this study due to the presence of 71 categories, some of which had fewer instances. This balance is crucial for ensuring that all categories are represented fairly in the classification process, as high accuracy alone can be misleading when dealing with imbalanced datasets. The time processing of each model was also considered.
5. **Iterative Labeling and Model Refinement:** After training the initial models on the manually labeled dataset, they were applied to datasets obtained from other supermarkets for a single day. An iterative process was employed to classify products from these supermarkets. The best-performing model from the initial evaluation was first used to classify products from the second supermarket, with its predictions reviewed and any inconsistencies manually corrected.

The corrected dataset from the second supermarket was then combined with the first to retrain the model, thereby enhancing its performance. This process was repeated for all six supermarkets, ensuring accurate ECOICOP category labels through expert review and correction. Ultimately, nearly 100,000 unique labeled products from all supermarkets were accumulated for this single day, allowing the final model to be retrained and used to label all products collected throughout 2022.

4. RESULTS

The results revealed that the CNN and BERTimbau models were the top performers for this classification task. The CNN model presented a robust accuracy of 96.8% and an impressive F1 Macro score of 87.3%, achieved in only 2.1 minutes. In comparison, the Portuguese-specific BERTimbau model achieved a slightly higher accuracy of 97.3%; however, its F1 Macro score of 72.2% lagged behind that of the CNN model. Furthermore, the processing time for BERTimbau was significantly longer, clocking in at 192.5 minutes. These findings suggest that while BERTimbau excels in accuracy, the CNN model provides a more balanced performance across categories, making it the preferred choice for short-text classification in this context.

With regard to large language models, BERTimbau significantly outperformed the others in terms of accuracy, achieving a rate of 97.3%, and for the F1 Macro score, obtaining a rate of 72.5%. Cross-language models such as BERT and XLM-RoBERTa, on the other hand, exhibited lower levels of accuracy, scoring 94.9% and 92.2% respectively, along with diminished F1 Macro scores (63.9% for

BERT and 58.9% for RoBERTa). The results clearly demon- strate the significance of language-specific large language models. BERTimbau achieved a remarkable F1 Macro score that was 12% higher than that of BERT, despite both models having the same architecture. This stark difference highlights the importance of language specific training in constructing highly efficient and effective language models.

5. CONTRIBUTION

This research contributes significantly to the field of official statistics and machine learn- ing in several ways:

• **Development of a Classification Pipeline:** This study successfully developed a complete machine learning pipeline for the automatic classification of food and beverage products using web-scraped data in Portugal. By transforming unstructured data into standardized classifications according to the European Classification of Individual Consumption According to Purpose (ECOICOP), the pipeline enhances the usability of web-scraped data for economic research and real-time monitoring of price indices.

• **Empirical Evaluation of Models:** The research provides a unique evaluation of various machine learning models for the classification of products in Portuguese, specifically highlighting the effectiveness of Convolutional Neural Networks (CNN) and the language-specific BERTimbau model. The findings illustrate the strengths and limitations of these models in the context of short-text classification, informing future applications in similar domains.

• **Foundation for Language-Specific Research:** The study confirms the value of language-specific models, like BERTimbau, for Portuguese text classification, providing a basis for future research into localized model development.

6. REFERENCES

[1] Aluko, V., & Sakr, S. (2019, December). Big SQL systems: An experimental evaluation. Cluster Computing, 22, 1347–1377. https://doi.org/10.1007/s10586-019-02914-4

[2] Aparicio, D., & Bertolotto, M. I. (2020). Forecasting inflation with online prices. International Journal of Forecasting, 36(2), 232–247. https://doi.org/10.1016/j.ijforecast.2019.04.018

[3] Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., & Trunfio, P. (2022, December). Programming big data analysis: Principles and solutions. Journal of Big Data, 9. https://doi.org/10.1186/s40537-021-00555-2

[4] Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation - Volume 6 (p. 10). USA: USENIX Association.

[5] Faridoon, A., & Imran, M. (2021, November). Big data storage tools using NoSQL databases and their applications in various domains: A systematic review. Computing and Informatics, 40(3), 489–521. https://doi.org/10.31577/cai_2021_3_489

[6] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[7] Harchaoui, T. M., & Janssen, R. V. (2018). How can big data enhance the timeliness of official statistics? The case of the U.S. Consumer Price Index. International Journal of Forecasting, 34(2), 225–234. https://doi.org/10.1016/j.ijforecast.2017.12.002

[8] Jahanshahi, H., Ozyegen, O., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2021). Text classification for predicting multi-level product categories. arXiv.https://doi.org/10.48550/ARXIV.2109.01084

[9] Lehmann, E., Simonyi, A., Henkel, L., & Franke, J. (2020, December). Bilingual transfer learning for online product classification. In Proceedings of Workshop on Natural Language Processing in E-commerce (pp. 21–31). Barcelona, Spain: Association for Computational Linguistics.

Retrieved from https://aclanthology.org/2020.ecomnlp-1.3
[10] Raasveldt, M., & Mühleisen, H. (2020). Data management for data science - towards embedded analytics. In Conference on Innovative Data Systems Research.
[11] Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In R. Cerri & R. C. Prati (Eds.), Intelligent Systems (pp. 403–417). Cham: Springer International Publishing.
[12] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... Stoica, I. (2016, October). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/293466