U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

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

Juliana de Freitas Ulisses Machado

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

PhD Bruno Miguel Delindro Veloso

PhD Miguel Reis Portela

2023

Abstract

The surge in internet usage has significantly amplified data generation, paving the way for researchers and policymakers to gain in-depth and granular insights about society. This shift has transformed data collection methodologies, fostering the exploration of non-traditional data sources such as web scraping. Numerous studies have demonstrated the efficacy of web-scraping in improving the accuracy of the Consumer Price Index calculation, outpacing traditional approaches due to its high frequency, granularity, and reliable nowcasting. However, this new paradigm presents two primary challenges: (1) devising appropriate frameworks for storing and processing large data volumes and (2) ensuring the quality of the collected data to facilitate its use by policymakers and researchers.

This research addresses these challenges in two phases. Initially, it delivers a comprehensive literature review on big data analytics, shedding light on state-of-the-art data storage and processing methods for large or unstructured datasets. Subsequently, it addresses a practical problem concerning the data acquired through web scraping from online grocery stores in Portugal for economic research. While the data collection process via scraping is relatively quick and cost-effective, complexity arises during the processing stage.

This study also explores machine learning techniques for short text classification, focusing primarily on Neural Networks and Natural Language Processing methods, to classify the collected products under the European Classification of Individual Consumption according to Purpose (ECOICOP) schema, the official classification used to calculate inflation in the European Union. It further extends its impact by proposing a method for the daily automated classification of products, utilizing the Convolutional Neural Network (CNN) model, which showed impressive performance with an accuracy of 96.60% and an F1 Macro score of 92.19%.

This project was carried out as part of an internship at the Banco de Portugal Microdata Research Laboratory within the European Master in Official Statistics (EMOS) program. It signifies a vital step in harnessing the potential of big data and machine learning to facilitate the production of comprehensive and timely official statistics.

Keywords: Big Data; Web Scraping; Natural Language Processing; European Classification of Individual Consumption according to Purpose (ECOICOP); Official Statistics; Short Text Classification; Product Classification; Convolutional Neural Network; BERTimbau

Resumo

A popularidade da internet resultou em um aumento substancial na geração de dados, o que representa uma oportunidade significativa para investigadores e para guiar políticas públicas, já que atualmente existe mais informação com uma maior granularidade para melhor compreender a sociedade. Este novo paradigma levou a uma transformação na abordagem adotada pelos institutos de estatística em relação à coleta de dados, levando à exploração de fontes de dados não-tradicionais, como web scraping. Por exemplo, vários estudos mostraram que extrair dados da web pode levar a resultados melhores para o cálculo do Índice de Preços ao Consumidor do que os métodos tradicionais. Além disso, pode antecipar mudanças na taxa oficial de inflação vários meses antes da abordagem tradicional devido à maior frequência e granularidade. Além das vantagens, este novo paradigma traz dois desafios críticos para os institutos de estatística: (1) o desenvolvimento da estrutura adequada para armazenar e processar grandes quantidades de dados e (2) a necessidade de garantir a qualidade dos dados coletados para permitir seu uso por investigadores e para fins oficiais.

Esta pesquisa pretende abordar estes dois tópicos. Primeiro, será fornecida uma revisão da literatura sobre big data analytics para apresentar o estado da arte sobre armazenamento e processamento de grandes conjuntos de dados e dados não estruturados. Posteriormente, aborda um problema prático relativo aos dados adquiridos por meio da raspagem de dados de supermercados online em Portugal. Embora o processo de coleta de dados por meio de raspagem seja relativamente rápido e econômico, a complexidade surge durante o estágio de processamento e uniformização desses dados.

Este estudo explora técnicas de aprendizado de máquina para classificação de textos curtos, com foco principalmente em Redes Neurais e métodos de Processamento de Linguagem Natural, para classificar os produtos coletados sob o esquema da Classificação Europeia do Consumo Individual de acordo com a Finalidade (ECOICOP), a classificação oficial usada para calcular a inflação na União Europeia. Além disso, é proposto um método para a classificação automática diária de produtos, através do modelo de Rede Neural Convolucional (CNN), que mostrou desempenho impressionante com uma precisão de 96.60% e uma pontuação F1 Macro de 92.19%.

Este projeto foi realizado como parte de um estágio no Laboratório de Pesquisa de Microdados do Banco de Portugal dentro do programa European Master in Official Statistics (EMOS), e representa um passo vital na utilização do potencial do big data e do aprendizado de máquina para facilitar a produção de estatísticas oficiais.

Palavras-chave: Big Data; Web Scraping; Processamento de Linguagem Natural; European Classification of Individual Consumption according to Purpose (ECOICOP); Estatísticas Oficiais; Classificação de textos curtos; Classificação de Produtos; Rede Neural Convolucional; BERTimbau

Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to Eurostat (2020a), the European Statistical System uses three main types of statistical processes to produce official statistics; the first is the direct collection of individual data from the Census, Probability Survey, and Non-Probability Survey. The second is the Administrative Data Process, where personal data acquired for administrative purposes, such as public sector accounting data, administrative register, and event-reporting systems, are used. Finally, the third one consists of two types of process: the Multisource Process and the Macro-aggregate Compilation Process, characterized by the combination of aggregated data from at least two sources to compile macro-aggregates in a specific field, such as inflation, consumption, and production.

These are the conventional data sources used by Statistical Officers, policymakers, and researchers. However, these data sources may have limitations in terms of measurement, such as slow reporting lag and lack of granularity. Nonetheless, it is well-established that the rise of the internet has contributed to the exponential growth of the available data. Therefore, Eurostat (2020a) also points out several pilot projects to test new techniques, including web scraping and text mining for generating statistics on online resources.

Singrodia et al. (2019) defines web scraping as the method to automatically obtain data from the URL of the website using coded programs in a structured approach. Web scraping can provide numerous advantages for official statistical officers, particularly regarding the collected data's extent, timeliness, and responsiveness. For instance, the ability to scrape prices from the web daily and report the results immediately allows for data collection that is not possible through manual means. Additionally, the flexibility to change the basket of goods in response to changing circumstances makes web scraping an ideal tool for official statistical officers. Finally, the cost and effort reduction of web scraping prices, as opposed to manually collecting them from physical stores, is also a significant advantage that needs to be considered.

The Billion Prices Project at the Massachusetts Institute of Technology was the first effort to generate a daily version of the Consumer Price Index (BPP CI) by scraping online retail prices of various products, as described by Cavallo & Rigobon (2016). The BPP CI provides real-time insight into the daily direction of consumer price inflation. In addition, they demonstrated that online price indexes could predict changes in official inflation rates months ahead. Finally, they state that collecting prices offline can be costly and complex, as it often requires many trained data collectors. In contrast, gathering prices online is less expensive, provides more detailed information about the product, and achieves high-frequency updates without delay.

Macias et al. (2022) performed a real-time nowcasting experiment using online food and non-alcoholic beverages prices obtained automatically from the most prominent online retailers in Poland. They demonstrated that only the estimation of online price changes was sufficient for predicting food inflation. Furthermore, they remark that the importance of online prices is increasing, especially after the COVID-19 outbreak, when the confinement measures led the physical stores to be closed, and e-commerce transactions rose rapidly. In conclusion, their results suggest that online prices can significantly aid central banks in their decision-making, given the crucial role of inflation nowcasts in macroeconomic projections.

In this rapidly-evolving context of emerging new data sources relying on large datasets of structured and unstructured data, it is essential to explore technologies suited to store, process, and efficiently analyze these datasets. Therefore, the primary goal of this research is to explore the potential of such technologies, presenting a literature review about big data analytics. In addition, the second objective of this work is to provide an official categorization for products obtained by web scraping from online grocery stores in Portugal, utilizing machine learning techniques to automate this task.

1.1 Motivation

The proposed work will be developed through an internship at Banco de Portugal Microdata Research Laboratory as part of the European Master in Official Statistics (EMOS) Program. Banco de Portugal Microdata Research Laboratory - BPLIM is an autonomous unit within the Economics and Research Department, with the core mission of supporting the production of research projects and studies about the Portuguese economy. It provides access to microdata sets customized to national and international researchers, who also can use the computational resources available at the laboratory. This research fulfils two critical demands of BPLIM: providing an in-depth review of big data analytics and investigating methodologies that can facilitate utilizing web-scraped data for research objectives.

1.2 Problem Description

The first part of this work provides a literature review of the leading technologies to process structured and unstructured datasets, comparing different types of Traditional and NoSQL databases and the main frameworks used to speed the processing of large datasets for data analysis and focusing on exploring the most used systems for distributed computing, such as Hadoop and Spark. BPLIM expressed interest in this literature review to understand better the required technology for managing large datasets in light of the current trend of exponential data growth and the emergence of new data sources.

Concerning the new data sources for economic research, there have been several initiatives to collect price data by web scraping since real-time data collection is a powerful tool for making near-term forecasts and increases granularity. However, Macias et al. (2022) highlights that acquiring scraped data is relatively quick and cost-effective, but adequately identifying goods that closely resemble the consumption basket requires a combination of machine and human classification, which is crucial for precise forecasting. Therefore, for over a year, Banco de Portugal has collected daily product and price information from online retailers. However, they are not using this data for research mainly because each product needs a standardized classification.

Hence, the second objective of this research is to explore Machine Learning techniques to predict multi-level product categories based on the product titles and brands to establish a data pipeline for classifying the products according to an official categorization to ensure that these databases can be used for research purposes. The classification system must be based on the European Classification of Individual Consumption According to Purpose (ECOICOP) published by the United Nations Statistics Division, United Nations Statistics Division (2018),and further subdivided by Eurostat to classify and analyze individual consumption expenditures. This classification is crucial to align with the European Union standards and is currently used to calculate inflation in the European Union. Eurostat (2020b) published a guide with practical guidelines on web scraping for official statistics purposes and identifies the automatic classification of product as the most critical challenge to process web scraped data efficiently. Nonetheless, the first hurdle this project must overcome is the absence of labelled data. Labelled data is crucial in training supervised machine learning models as it allows them to understand the relationships between input and output variables, effectively serving as a roadmap for the model to make precise predictions or classifications. Consequently, an essential part of this project also revolves around generating the first set of labelled data. This dataset will serve as the foundational training material for the models, essentially kick-starting the automated classification process.

1.3 Thesis Structure

This thesis is organized as follows:

Chapter 2 provides the Literature Review. First, an in-depth examination of Big Data is provided, emphasizing the various storage technologies suitable for structured and unstructured data. These technologies' main advantages and limitations are discussed concerning specific use cases, and a comparative analysis of SQL-based systems and NoSQL is presented. Furthermore, an overview of the most commonly used file formats for analytical tasks and an introduction to the Apache Arrow Project are provided. Then an overview of distributed computing frameworks, such as Hadoop, Hive, Impala, and Apache Spark, is presented. Towards the end of Chapter 2, the spotlight is turned toward using web scraping for official statistics. The chapter also explores various machine-learning techniques that can be employed to address the intricate problem of product classification.

In Chapter 3, the focus is shifted to the practical aspects of the research. This chapter elucidates the data, software, and methodology deployed in this project, providing an operational backdrop against which the subsequent results can be understood.

Chapter 4 then presents the research questions and the outcome of the investigative processes outlined in Chapter 3. This section provides a detailed discussion of the results obtained, providing an in-depth analysis of this task.

Finally, Chapter 5 ties the entire research together by offering a comprehensive conclusion. It reflects on the overall project, emphasizing key findings and their implications. It also outlines potential avenues for future research, thereby providing a roadmap for subsequent studies in this area.

Chapter 2

Related Work

This chapter offers a comprehensive literature review of the primary paradigms in Big Data Analytics, detailing storage systems, file formats, and distributed computing. The structure of the first section is as follows:

First, storage technologies applicable for structured and unstructured data are discussed, alongside a presentation of the four major categories of NoSQL databases. This is followed by comparing traditional database systems and NoSQL storage systems. In addition, there's an introduction to NewSQL databases and embedded OLAP database management systems.

The subsequent subsection delves into the various formats for data storage, such as CSV, JSON, ORC, Parquet, Avro, FST, and Feather. This section also briefly assesses their performance and introduces the Apache Arrow Project.

The third subsection focuses on the primary bottlenecks encountered when processing large datasets, presents the notion of distributed computing and parallel processing, and surveys the main frameworks for distributed computing like Hadoop, Spark, Hive, and Impala.

To conclude, a literature review about applying web-scraped data in official statistics is presented, which includes methodologies for short text classification models, particularly emphasizing product categorization.

2.1 Big Data Analytics

Gandomi & Haider (2015) explores a broader definition of big data combining concepts from business professionals and researchers, highlighting that the data dimension is only one property considered in big data. They point out that the paradigm of large-scale data has three major challenges concerning volume, velocity, and variety, referred to as the 3V model for big data. Volume refers to the total size of the datasets, velocity pertains to the pace at which data is acquired and subsequently prepared for analytical purposes, and variety encompasses the diversity of data representation, including structured, semi-structured, and unstructured data. However, this model has extended to include other challenges, such as veracity, value, and variability. Veracity refers to the inaccuracies, noise, and irregularities in data; value is the characteristic of data that pertains to how it can generate value from a business viewpoint; and variability refers to the dynamic changes in the data flow rate. Therefore, this section will present data storage and processing technologies according to the broader definition of big data.

2.1.1 Storage Technologies

Big data storage is an infrastructure designed for storing, processing, and retrieving vast data. Faridoon & Imran (2021) provides a systematic literature review from 2015 to 2020 by analyzing recent publications about NoSQL storage systems. The authors discussed the motivation and necessity for new technologies for big data storage. They argued that relational, structured, and well-schema databases struggle to keep pace with big data's rapid growth and complexity. Also, the large volume of data often needs real-time processing, fast recovery, fault tolerance, and complex data structure. Overall, management and dynamic scalability requirements exceed the capabilities of relational databases. Table 2.1 compares the strengths, weaknesses, opportunities, and threats of Traditional Databases Systems and NoSQL Storage Systems.

Table 2.1: Traditional Databases Systems and NoSQL Storage Systems - Adapted from Faridoon & Imran (2021)

	Traditional Database Systems	NoSQL Storage Systems
Strengths	Store structured data under a predefined schema which provides consistency in data. This structure enables efficient data management. It ensures a high level of data integrity through the use of transactions. Vertical scalability.	Store un-structured, semi-structured, and structured data. Horizontal scalability and supports parallel computing. High availability and fault tolerance. Reliability. Simultaneous accessibility and consistency
Weaknesses	Bottleneck performance when data increases. Limited storage capacity. Difficulty in join operations for multidimensional data.	No support for ACID (Atomicity, Consistency, Isolation, Durability) properties.
Opportunities	RDBMS support complex queries. Consistence in complex database transactions.	Simplicity in complex storage structures. Speed time for query.
Threats	As data increases, large volume for storage is required. Complex and schema less data structures. Real-time processing and maintaining consistency as storage servers increases.	Deployment. Difficulties while processing small files in large numbers.

NoSQL is flexible and does not require a predefined schema structure; it provides the flexibility to store entire data in terms of documents instead of the traditional method of table-row-column. NoSQL is extensively useful when accessing and analyzing vast amounts of unstructured data or data stored remotely on multiple virtual servers Tauro et al. (2013).

Also, unlike relational databases, NoSQL does not support ACID properties, such as Atomicity, Consistency, Isolation, and Durability, that ensure secure data storage during database transactions. The ACID properties affect the performance of the relational databases, and in some cases, the speed of the transactions is the priority. In these situations, the BASE properties such as Basic Availability, Soft-state, and Eventual consistency are more appropriate and are used in NoSQL databases. Therefore, NoSQL databases and relational databases are designed for different purposes; relational databases are more stable and reliable over time and have more support from professional services and companies, while most NoSQL databases are open-source tools most suited for speed while processing large amounts of structured, semi-structured and unstructured data by different levels of reliability of available data, as explored by Uyanga et al. (2021).

There are four major categories of NoSQL databases:

- Key-value stores: The data are stored as key-value pairs. This data storage supports structured and unstructured data and can be easily deployed in a distributed environment. They are very beneficial for high-speed reading and writing of non-transactional data. Stored data is partitioned and further replicated to achieve scalability and availability. Some examples of key-value databases are Redis, Aerospike, and Riak, which are commonly used in web applications such as shopping cart data storage.

- Columnar-stores: In this data store, the scalability is reached by splitting rows and columns over multiple nodes. This system provides efficient data compression and partition and performs well with basic aggregation queries. Also, this system is highly scalable and allows parallel processing. Some examples of columnar stores are Hbase, Cassandra, Hypertable, and Bigtable. The preferred areas are blogging platforms and content management systems; Linkedin uses HBase, Cassandra is widely used for online interactive applications like Facebook and Twitter, and Google created BigTable to store web pages and other products.

- Document-based stores: In this system, data is stored and organized as document collections instead of structured tables. It supports secondary indexes and can be easily updated with new fields and records. It was implemented in SimpleDB, CouchDB, and MongoDB.

SimpleDB is used for complex queries, logs, and online games; CouchDB is often used for web applications and social data; MongoDB is the most used NoSQL database; it is commonly used for real-time applications.

- Graph databases: These databases use nodes, edges, and properties to represent relationships among data in graphs. This system is implemented in Neo4j, InfiniteGraph, HyperGraphDB, etc. Neo4j is used for social networks and recommendation systems, and HyperDB is most used for bioinformatics, pattern mining, and semantic web projects. Infinitegraph has been used for social and location-based networks and real-time searches.

Table 2.2 provides an overview of the storage systems previously cited.

Table 2.2: Properties of the most used NoSQL systems. Adapted from Faridoon & Imran (2021)

	Storage Systems	Memory Storage	Disk Storage	Intensive Read/Write	Persistance	Partitioning	Shared nothing Architecture	Scalability
Key-value	Aerospike	✓	X	✓	X	✓	✓	✓
	Rick	✓	✓	X	✓	X	✓	X
	Redis	✓	X	✓	✓	✓	✓	✓
Columnar	Hbase	X	✓	X	✓	✓	X	✓
	Cassandra	✓	X	✓	X	✓	✓	✓
	Hypertable	✓	✓	✓	✓	✓	✓	✓
	BigTable	✓	✓	✓	✓	✓	✓	✓
Document-based	MongoDB	✓	✓	✓	X	✓	✓	✓
	SimpleDB	X	X	X	✓	X	X	✓
	CouchDB	X	✓	✓	✓	✓	✓	✓
Graph databases	Neo4j	✓	✓	✓	✓	X	✓	✓
	HyperGraphDB	✓	✓	-	X	✓	-	✓
	InfiniteGraph	-	-	✓	X	✓	✓	✓

There have been many efforts in the last decade concerning the speed performance of SQL and NoSQL Databases. Tauro et al. (2013) stated that conventional relational database systems are not effective for huge data queries and highlighted that scalability in relational databases requires powerful servers that are expensive and complex to manage. Also, it remarks that it is expected to use NoSQL databases when they are designed to scale as they grow.

Kausar & Nasar (2021) also explored the rise of NoSQL for Big Data Analytics. The authors evaluated the performance of NoSQL and traditional SQL and concluded that the NoSQL database is a superior alternative for industries that need high-performance analytics, adaptability, simplicity, and horizontal scalability.

R. Wang & Yang (2017) also tested the read and write speeds of the relational database MySQL and two types of NoSQL databases, key-value storage (BerkeleyDB) and document storage (MongoDB), in a single-machine environment. NoSQL was much faster than traditional SQL databases for reading, especially the key-value storage. Concerning writing tasks, BerkeleyDB was considerably quicker than MySQL and MongoDB. However, MySQL became faster than MongoDB when data collection increased to 1,000,000 records.

Finally, Kanchan et al. (2021) provided extensive experiments to measure the performance of MySQL, MongoDB, Cassandra, and Redis for execution time and memory consumption for moderately sized and structured datasets. They concluded that NoSQL outperforms SQL-based systems in basic read and write operations. However, SQL-based systems are better if queries on the dataset mainly involve aggregation operations. Overall, the most efficient database for write, update, and delete operations was Redis; MongoDB gives the worst performance when the size of data increases but shows the best performance for efficient memory consumption for reading tasks. Concerning performance, Rides outperformed Cassandra, MongoDB and MySQL. MySQL provides better performance for aggregate functions, and NoSQL was unsuitable for complex queries and aggregate operations.

Overall, relational databases have been traditionally employed for structured data where transactional consistency is a crucial aspect of the system, as in the case of financial systems, e-commerce platforms, and enterprise resource planning (ERP) systems. On the other hand, NoSQL databases are well suited for unstructured or semi-structured data, and scalability is a vital aspect of the system. These databases are particularly suitable for applications that handle large volumes of data, have a high data ingestion rate, and are designed for horizontal scalability. Examples of such systems include social media platforms, gaming systems, and real-time analytics.

As mentioned, the recent exponential growth in the volume and complexity of generated and stored data has necessitated the development of NoSQL databases specifically designed for scalability. However, these systems often require more consistency and transactional guarantees. In this context, a new class of database management systems called NewSQL has emerged, attempting to provide the scalability and performance of NoSQL databases while maintaining the consistency and transactional guarantees of traditional RDBMS by maintaining ACID guarantees for transactions. However, NewSQL systems have had a relatively slow adoption rate since the leading DBMS vendors, such as Microsoft, Oracle, and IBM, are well-established. Also, these prominent vendors choose to innovate and improve their DBMS systems rather than acquire NewSQL start-ups. Pavlo & Aslett (2016)

Concerning the performance of NewSQL databases for online transaction processing for Big data management, Kaur & Sachdeva (2017) provided experiments to compare the most prominent solutions in this technology, namely NuoDB, VoltDB, MemSQL, and Cockroach DB regarding the read latency, write latency, update latency, and execution time. According to the results, NuoDB performed better in most test cases, followed by MemSQL, which had the best performance in updation latency. CockroachDB, on the other hand, was the slowest system for almost all operations except execution time, where VoltDB had the worst performance.

Beyond the necessity for data storage, the choice of storage technology directly affects the processing time of analytical tasks. Data Science has led to a significant increase in the complexity of data analysis, making advanced tools beyond standard SQL queries necessary. Raasveldt & Mühleisen (2020) claims that data analysts usually use a combination of independent solutions for data storage, utilizing scripting languages like R and Python on individual computers rather than relying on traditional RDBMS and large dedicated servers.

From that motivation, Raasveldt & Mühleisen (2019) presented what they called ”a new class of data management systems” designed to work as an embeddable analytical data management: DuckDB, an open-source data management system designed to execute analytical SQL queries in a fast and efficient way, while embedded in another program, such as R and Python.

The authors invoke the importance of embedded systems, citing the example of SQLite, an embedded system that is the most widely deployed SQL database engine. Nevertheless, they highlight that SQLite is an OLTP system designed for transactional workloads and performs poorly on analytical tasks. It highlights the latent need for embeddable analytical data management, mainly for data analysis, that provides an API for R and Python, where packages like dplyr and Pandas are commonly used but do not offer query optimization and transactional storage.

They also state that another advantage of embedded analytical data management is related to edge computing scenarios, meaning that the data will be processed closer to the source. Consequently, there is less need to transmit large amounts of data to a central location, reducing bandwidth usage and increasing security protection by keeping sensitive information within the secure perimeter of the network.

2.1.2 File Formats and Apache Arrow

Another important aspect concerning processing a large amount of data is its format; this influence directly the volume and the speed of analytical tasks. The space required to store the data directly affects its local or cloud storage costs, and each data file type can have its compression scheme. Also, the encoding scheme of the file format affects performance according to the required task. The most commonly used file formats for data analytics are listed below, and a comparison is provided.

- The Comma-Separated Values (CSV) is a text-based file format generally delimited by the comma character; however, it is possible to use separators other than commas, such as tabs or spaces. CSV is a row-based file format; each row represents a separate data record. Generally, the first row contains column headers, and there is no support for column types. It is a popular file to store tabular data and is commonly used by spreadsheets programs like Excel and Google Sheets. This file format is simple to read and write, can be used in most data analysis tools, can be read using almost any text editor, and is compact compared to other row-oriented data files, such as XML and JSON. Nevertheless, CSVs do not provide scheme support and require that complex data structures be processed separately from the format, increasing the processing time while handling large datasets.

- JavaScript Object Notation (JSON) is presented as key-value pairs in a partially structured format, supporting hierarchical structures that simplify storing related data in a single document and presenting complex relationships. JSON files are human-readable and allow data to be stored in many types, such as strings, integers, objects, arrays, Booleans, and null. It is a standard format for storing textual data based on JavaScript object index, been widely used for transmitting data in web applications. It is easy to integrate with APIs, can hold vast data efficiently, and has been commonly used for non-tabular databases. JSON files are more compact than XML files, although it consumes more memory than CSV due to repeatable column names. Also, it is less compact than binary formats.

- Avro is an open-source data serialization system that helps exchange data between systems, programming languages, and processing frameworks. Avro can define a binary format for the data and map it to many programming languages. The schema is stored in JSON format, while the data is stored in binary format, which minimizes file size and maximizes efficiency. Therefore, it is very efficient for storing row data, supports file splitting, and has excellent integration with Apache Kafka. It works best to write data, and it is slower to read. This data file is the leading serialization format for record data and is the first choice for streaming data pipelines.

- Feather V2 / Arrow IPC is a lightweight binary format that utilizes the Arrow IPC format internally for storing data frames by the Apache Arrow Project. It was designed to do reading and write data frames efficiently and to make sharing data across data analysis languages easy. It supports several column types formats. Feather was created as a proof of concept for fast, language-agnostic data frame storage for Python (pandas) and R. It supports compression with LZ4 and ZSTD. Also, it provides memory mapping that improves performance and can minimize memory overhead for processing.

- Apache Parquet is a popular open-source column-oriented data file format created to make compressed, efficient columnar data representation available to any project in the Hadoop ecosystem. Parquet stores data in a binary format that provides highly efficient data compression and decompression and has been widely used to store big data files such as structured data tables, images, videos, and documents. Also, Parquet supports nested data and schema evolution. It is the default data source in Apache Spark and is available in multiple languages such as Java, C++, Python, R, etc.

- Optimized Row Columnar (ORC) is an open-source column-oriented data storage format that provides a highly efficient way to store Hive data. It has efficient compression since it uses compressed column storage, which leads to smaller disk reads. Also, ORC has a built-in index, min/max values, and other aggregates that cause entire stripes to be skipped during reads; this makes this data file fast to read. It is proven in large-scale deployments like Facebook, which use the ORC file format for a 300+ PB deployment.

- FST is a binary columnar format created for R that offers a quick, convenient, and adaptable method to serialize data frames. The FST format supports full random access to columns and rows in stored data frames. It heavily relies on multi-threading for improved read and write speeds. Klik (2022).

Columnar file formats store structured data in a column-oriented way. The main advantage of this approach is that it provides an efficient way to read only a subset of columns, which is very efficient in optimizing queries. Also, data organized by columns provide homogeneity, which supports better encoding and compression to achieve better speed and file size efficiency.

It was observed that the format selection and its configuration drastically affect the performance. Columnar file formats like Apache ORC and Apache Parquet are potent tools for optimizing query execution. Indeed, ORC performs better on Hive, while Parquet is the best choice with Spark SQL, as analyzed byIvanov & Pergolesi (2020). Also, according to Gohil et al. (2022), ORC is the more efficient while dealing with integer and string data types in terms of space, whereas Parquet is the best choice for decimal-type data.

In an extensive experimental study for scalable query execution engines for massive structured data, Aluko & Sakr (2019) found that the Parquet file format provided the highest performance for most queries. Nevertheless, they stated that converting data from the Textfile to the Parquet format was extremely time-consuming with large datasets. Additionally, memory utilization became very high without an adequately defined partition for the Parquet file.

Regarding Feather (Arrow IPC) file format and the Parquet format, the The Apache Software Foundation (2019) claims that Parquet is a storage format optimized for maximum space efficiency through advanced compression and encoding techniques for long-term storage and archival use. At the same time, the Arrow on-disk does not prioritize the requirements of long-term archival storage. Also, it is essential to highlight that parquet files are much smaller than Arrow IPC files because of compression strategies; this implies choosing parquet files if disk storage or the network is slow, even for short-term storage or caching. Lastly, reading parquet files requires complex decoding. At the same time, Arrow data is not compressed, meaning it is an in-memory format meant for direct and efficient use for computational purposes. Therefore, storing the data on the disk using Parquet and reading it into memory in the Arrow format is the best scenario for optimizing computing hardware.

To conclude the performance comparison, Ursa Labs (2019), an industry-funded development group specializing in open-source data science tool to help make Apache Arrow a robust and reliable next-generation computational foundation for in-memory analytics, have tested columnar file performance for Python and R using Parquet, Feather, and FST files. Parquet was recommended as a gold-standard column file format. The benchmarks showed that the performance of reading the Parquet format was similar to other columnar structures but came with additional benefits like the size of Parquet files due to Parquet's compression schemes, and this is the industry-standard format for data warehousing since it is compatible with Apache Spark and nearly any modern analytic SQL engine.

Concerning file formats, as mentioned before, Feather was created as a language-agnostic data frame storage for the Apache Arrow Project. However, Apache Arrow goes far beyond Feather, considered only a proof of concept for this project. Apache Arrow provides an in-memory columnar data format that efficiently transfers data between systems. The main objective of this approach is to accelerate big data workloads and improve reading performance. It also provides interfaces for different languages, making possible zero-copy inter-process communication.

Topol (2022) states that the primary goal of Arrow is to become the lingua franca of data analytics, eliminating the need for different internal formats for managing data since moving data between these components has a considerable cost to serialize and deserialize every time. In addition, Arrow focuses on an in-memory format, aiming for CPU efficiency through tactics such as cache locality and vectorization of computation, which are not possible with on-disk formats. Also, data stored on disk poses challenges regarding size, and input/output (I/O) costs when reading it, which is much slower than memory access.

Several projects have been using Arrow in their internal and external communication formats. For example, concerning analytical tasks, Polars, a lightning-fast DataFrame library/in-memory query engine for Python, provides a highly efficient API for data wrangling and data pipelines built upon the Apache Arrow. In addition, DuckDB can be integrated with Apache Arrow to analyze larger-than-memory datasets directly in Python and R rapidly. Also, many projects use the Arrow JavaScript library to perform super-fast computations inside the browser.

2.1.3 Distributed Computing

Khan (2015) defines distributed computing as the practice of utilizing multiple computers connected by a network in order to share and divide a workload. Distributed computing aims to distribute the workload among multiple machines to achieve more efficient and faster processing. In order to maintain the availability of the service in the event of component failures, the system must be designed with redundancy in both space and time, pointing out the importance of the fault-tolerant property in distributed computing since the greater the tolerance for failures, the higher the level of resilience and dependability in the distributed system as a whole. The scalability of a system can be reached by adding more computers to the cluster.

On the other hand, parallel processing is a programming paradigm in which a computation is divided into smaller units that can be executed concurrently on multiple processors in a single computer. This approach can improve performance since it allows multiple processors to work on a problem simultaneously, significantly reducing the time it takes to complete a computation. The scalability of a system can be reached by adding more processors for the computations, and the practice of managing the workload across multiple processors can lead to better management of the available resources.

Distributed and parallel computing approaches can handle the main bottlenecks that can arise when processing large datasets, for instance:

- CPU (Central Processing Unit): The CPU is the main processing unit of a computer and is responsible for executing instructions and performing calculations. If the CPU is not powerful enough or overloaded, it can become a bottleneck in data analysis.
- RAM (Random Access Memory): It is a volatile memory that temporarily stores data while a computer is running. If the amount of data being processed exceeds the capacity of the RAM, it can lead to slower processing speeds and bottlenecks.

- Disk I/O: Disk input/output refers to the speed at which data can be read from and written to a disk. As the volume of data being processed increases, the disk may become a bottleneck if it is not fast enough to keep up with the rate at which data is being processed.
- Network I/O: If the data being analyzed is stored on a remote server or accessed over a network, the speed and reliability of the network can impact the performance of the data analysis process.

Dean & Ghemawat (2004) proposed the MapReduce framework as a powerful interface to enable automatic parallelization and distributed large-scale computations to process large amounts of raw data at Google in 2004. Since then, it has become a popular paradigm for distributed computing. The MapReduce model consists of two main functions: the map function and the reduce function. The map function processes a key/value pair to generate a set of intermediate key/value pairs. The reduce function merges all intermediate values associated with the same intermediate key. Another central aspect of this framework is fault tolerance through redundant execution of tasks and periodic checkpointing of the computation state. This mechanism enables MapReduce to continue the computation despite failures, ensuring reliable and efficient processing of large data sets. Furthermore, MapReduce is a batch processing system since it can automate and process multiple transactions as a single group.

The most famous open-source software implementing the MapReduce programming model is Apache Hadoop, which links the MapReduce processing engine with a distributed file system, the Hadoop Distributed File System (HDFS). Shvachko et al. (2010) presented the Hadoop Distributed File System (HDFS) architecture and described the experience of using HDFS to manage petabytes of data at Yahoo!. Concerning the architecture, HDFS stores file system metadata and application data separately. The NameNode dedicated server stores metadata, and the other servers, called DataNodes, store the application data. Multiple DataNodes also contain a replication of the file content to ensure reliability. Furthermore, preserving server metadata by HDFS enables the provision of an API that comprises the coordinates of file blocks. It enables the MapReduce framework to arrange tasks close to the data, enhancing the I/O bandwidth and thus augmenting the efficiency of data retrieval.

Overall, Hadoop addresses the challenges of expanding storage and computational capacity by leveraging a cluster-based architecture and distributing data across multiple nodes in a network. This storage system enhances processing power and supports scalability by adding more nodes to the cluster. In addition, by distributing the storage and processing of data across multiple nodes, Hadoop significantly reduces the time required for handling large volumes of data.

Vavilapalli et al. (2013) introduces the new architecture for Hadoop, the YARN (Yet Another Resource Negotiator), by presenting a massive-scale production experience of Yahoo!. YARN is a resource management system that addresses some of the limitations of the original Hadoop MapReduce framework with the initial goal of improving scalability in Yahoo! workload. However, as reported in the article, it considerably improved resource utilization by allowing different workloads to share the same cluster and even removed the need to scale further. Also, YARN provided more flexibility allowing different data processing frameworks, such as Spark, Hive, and Storm, to run on top of Hadoop, giving users more options for working with their data. In addition to MapReduce, HDFS, and YARN, Hadoop has a Common module with a set of standard utilities needed by the other modules, such as implementations for compression codecs, I/O utilities, and error detection.

According to White (2015), the term ”Hadoop” has gained another dimension beyond HDFS and MapReduce since it is commonly used to describe a larger ecosystem of projects related to distributed computing and large-scale data processing. Some of the related projects of Apache Hadoop are:

- Apache Flume: A distributed and reliable service for collecting, aggregating, and moving large amounts of event-based data, like log data, into Hadoop.

- Apache Sqoop: A tool that transfers data between Hadoop and relational databases. It allows users to import data from structured data stores such as relational databases, data warehouses, and enterprise data systems into the Hadoop Distributed File System (HDFS).

- Apache Pig: A platform to facilitate the development of Big Data analysis applications. It provides a data flow programming language called Pig Latin to execute MapReduce jobs using SQL-like syntax. It allows programmers with a good SQL background to process large amounts of unstructured data using the Hadoop framework.

- Apache Hive: A framework for data warehousing on top of Hadoop built by Facebook. Hive enables analysts with advanced SQL proficiency but limited Java programming abilities to execute queries on large amounts of data stored in HDFS. It provides a simple, SQL-like language called HiveQL to perform data analysis and querying on large datasets. Hive translates the HiveQL queries into a series of MapReduce jobs, enabling analytics at a massive scale using the Hadoop cluster.

- Apache Impala: It is an open-source, massively parallel processing SQL query engine designed to provide users with a familiar SQL interface while taking advantage of the scalabil-

> ity and flexibility of Hadoop. Impala combines the functionality of a traditional analytical database with the ability to handle big data and multi-user performance in a Hadoop environment. Kornacker et al. (2015)

Beyond Apache Hadoop, another open-source, distributed computing system designed to process large data sets quickly is Apache Spark, developed at the University of California, Berkeley. Zaharia et al. (2010) introduced Spark as a cluster computing framework that employs a shared data set across multiple parallel operations while maintaining comparable scalability and fault tolerance characteristics to those of the MapReduce framework with a data-sharing abstraction called Resilient Distributed Datasets or RDDs. The article reported that Spark outperformed Hadoop by 10x in iterative machine learning workloads.

In 2016, Zaharia et al. (2016) pointed out that since its launch in 2010, Spark had become the most active open-source big data processing project, with over a thousand contributors and numerous users across various industries. This significant increase in participation has driven efforts to make Spark a unified engine. Many contributors have worked to create an integrated standard library that covers everything from data import to machine learning, intending to address a wide range of processing requirements previously required by a separate engine.

The main benefits of Spark are as follows: the use of a unified API simplifies the development of applications, the ability of Spark to perform various functions on the same data in memory allows for a more efficient combination of processing tasks, and Spark also enables the implementation of new types of applications, including interactive queries on graphs and streaming machine learning, which was previously unreachable with prior systems.

Ismail et al. (2019) conducted experiments to evaluate the performance of Apache Spark and Hadoop for machine learning tasks in terms of performance, storage, reliability, and architecture. The study compared Impala and Hive as query engines on the Hadoop file system against the native Spark query engine. The results showed that Impala performed better than Hive and Spark in small-scale workloads, with better average performance. In addition, Impala performed better in moderate memory capacity scenarios than the other platforms, while Spark was the fastest in high-memory environments. Hive stood out when there was a lack of memory, while the other platforms failed in the same conditions. Overall, Spark outperformed Hadoop MapReduce on complex queries and performed well with any volume of data when there were no limitations on memory consumption. However, Hadoop MapReduce was the better choice when faced with limited memory space and a need for speed.

Aluko & Sakr (2019) also provided an extensive experimental study of four popular systems for scalable query execution engines for processing massive structured data while supporting SQL interfaces, namely Apache Hive, SparkSQL, Apache Impala, and PrestoDB. They analyzed these systems' performance using three benchmarks for several analytical tasks. In the experiments performed on the 99 queries of the TPC-DS benchmark, on average, Apache Impala has the fastest query execution times, while Hive is the slowest, and Spark SQL and Presto have shown a comparable good performance. In all benchmarks, Spark significantly outperformed Hive.

Belcastro et al. (2022) reviewed parallel and distributed approaches, languages, and systems for analyzing and processing Big Data on scalable computers. They gave an overview of the critical features of prominent distributed computing frameworks. They also discussed the most commonly used Big Data analysis systems and compared these systems, highlighting their main features and the pros and cons of each for implementing Big Data applications. A summary of the advantages and disadvantages of these systems is shown in Table 2.3.

Table 2.3: Advantages and disadvantages of distributed computing frameworks. Adapted from Belcastro et al. (2022)

System	Advantages	Disadvantages
Hadoop	Fault tolerance, low cost, very large open source community	Complex coding, only support batch processing, small files issues, inefficiency with iterative applications
Spark	In-memory computing, easy to code, flexibility, libraries for advanced analytics, scalable machine learning support	No automatic optimization process, small files issues, high memory consumption
Storm	Multi-language support, low-latency response time	Message ordering not guaranteed
Hama	Many Distributed FS-supported, general-purpose computing on GPUs, conflicts and deadlines avoidance	Single point of failure (BSP Master), low flexibility of partitioning policies, small community
Hive	Large distributed datasets querying, SQL-like language, UDFs for advanced data analysis	Support only for OLAP, real-time data access not supported
Pig	High-level procedural language, UDFs for advanced data analysis	Small community, hard to tune performance

Regarding the verbosity, Belcastro et al. (2022) classified each system as follows:

- High: Systems that require complex programming instructions for simple tasks. For example, writing a MapReduce task in Hadoop;

- Medium: These systems necessitate the implementation of specific interfaces and methods to encode an application. Some systems in this category are Storm and Hama.

- Low: Systems that require minimal coding, often offering a user-friendly programming approach. Some examples are Spark, Hive, and Pig.

The previous table mentions two new software programs: Apache Hama and Apache Storm. Apache Hama is an open-source framework built on the BSP model, optimized for handling complex tasks such as matrix and graph computation on small-scale infrastructure. It mainly supports graph processing applications like graph analysis, deep learning, and machine learning, leveraging the BSP model. Apache Storm is an open-source, real-time stream processing system for large-scale infrastructure handling massive unbounded data. It is designed for high scalability, fault tolerance, and fast data processing and can process millions of tuples per second per node with low latency.

Landset et al. (2015) explores the Hadoop ecosystem, presents the advantages and disadvantages of different processing engines, and compares them, focusing on implementing machine learning techniques. They provide an overview of the current open-source, scalable machine learning tools. The study concluded that MapReduce is becoming outdated and is not recommended for most applications due to its slow performance and lack of support for iterative algorithms. They remark that Spark is a more suitable alternative. For real-time solutions, Storm and Flink were suggested as alternatives to Spark Streaming as they offer more accurate stream processing, unlike Spark's use of micro-batch streaming which can cause a slight delay in receiving results. Flink offers a combination of batch and true stream processing, but it is a relatively new project and requires more research on its viability. Additionally, it has limited support for machine learning solutions compared to other platforms. H2O was noted as the exclusive complete system in the study and provided two distinctive features, such as a user-friendly graphical interface and support for deep learning. It rivals, if not surpasses, other machine learning platforms in terms of the tools it supports. Despite this, further research is required to assess H2O fully. Mahout and MLlib were acknowledged as the most comprehensive big data libraries regarding the number of algorithms they cover and their compatibility with Spark and H2O.

Concerning using ”big data” for social science research, especially economics and finance, Bluhm & Cutura (2022) provides a valuable guide for economists who want to analyze datasets larger than their computer's memory allows. The article includes codes and configuration instructions for examples featuring various micro-, panel- and time-series econometrics applications using Spark that do not require a background in computer science and parallel/distributed computing or having physical access to high-performance computers.

They highlight that despite Python and R being inferior to higher-level programming languages like C++ and Julia, their experiments indicate that for empirical economic research, Python and R, combined with Spark, are suitable tools for economic research when handling and analyzing massive datasets.

2.2 Multi-level Product Classification of Web Scraped Data

This section presents the use cases of web-scraped data for official statistics and the methods that will be explored for the analytical problem of multi-level classification of products obtained via web scraping from online grocery stores in Portugal.

2.2.1 Web Scraping and Official Statistics

In the U.S., the principal agency for the Federal Government in the field of labour economics and statistics aims to modernize the Consumer Price Index data collection. The Bureau of Labor Statistics (BLS) states that one of its objectives is to convert a significant proportion of the CPI market basket from traditional collection to nontraditional sources, including utilizing large-scale data, by 2024. National Academies of Sciences, Engineering, and Medicine (2022).

Harchaoui & Janssen (2018) explores the use of big data to enhance the timeliness of official statistics by using the Billion Price Products CPI (BPP CPI) and discusses the U.S. consumer price index case. They discuss that the ”Billion Price Project” showed that the web is becoming a considerable alternative for price collection and can lead to cost and time reduction when processing official statistics regarding inflation. In addition, they claim that the daily BPP CPI can offer a more reliable estimate of a timely official CPI under a suitable modelling strategy.

Aparicio & Bertolotto (2020) used online price indices to forecast the Consumer Price Index (CPI). The results of their study showed that their approach outperforms the most common benchmarks in the literature and two leading surveys of professional forecasters. This is mainly due to the high frequency of data available from online prices, which allows for more accurate predictions. The study also found that their baseline one-month forecast outperformed statistical benchmark forecasts for Australia, Canada, France, Germany, Greece, Ireland, Italy, the Netherlands, the United Kingdom, and the United States.

Hillen (2019) discussed web scraping as a method for extracting large amounts of data from online sources for food price research. The research showed that web scraping is a promising method to collect customized, high-frequency data in real-time, overcoming several limitations of currently used food price data sources. Some of the advantages of this technique are the low cost, high frequency, product details, store aggregation, transparency, and customization of the dataset according to the research purposes. However, some of the limitations of this technique include ethical and legal uncertainties, the absence of historical and transaction data, the need for online availability, and technical difficulties in handling the amount of information collected.

Mahajan & Tomar (2021) analyzed the impact of lockdowns provoked by COVID-19 outbreak in the food supply chains in India by using web-scraped data from a major online grocery store in India to examine the effect on product shortages and pricing. The research found that online product availability fell about 10 per cent, and there was little effect on the prices of goods sold online. They highlighted that products made far from retail centres faced higher shortages, revealing the vulnerability of supply chains for items that travel long distances before reaching the final point of sale. This study demonstrated the importance of online data for timely decision-making, especially during critical situations like the COVID-19 outbreak when official data collection was affected.

An important issue concerning this technique is whether online prices are comparable to those in physical stores. A study by Cavallo (2017) compared prices from the websites and physical stores of 56 large retailers across ten countries and found that prices were identical about 72 per cent of the time. The study found that while price changes were not always synchronized, they had similar frequencies and average sizes. Therefore, they concluded that online prices could be considered a reliable source for retail prices for Statistical Officers for consumer price indexes as an alternative data-collection technology for multichannel retailers.

Several studies show the importance of web-scraped data. Furthermore, as the availability of price data on the internet increases, the scope of web-scraping applications will also expand and can be extended to other fields. This scenario reinforces the importance of developing the right tools for automatizing product classification according to official categories to allow web-scraped data to be a valuable resource for research purposes and to produce official statistics.

2.2.2 Multi-level Product Classification

Bertolotto (2016) uses data gathered from online sources for economic research and describe the adopted methodology for the product classification schema. They use a system that applies machine learning techniques to automatically suggests product categorization (COICOP), packaging size, measurement unit, and whether it is a single item or a bundle. Then, a team of trained specialists evaluates these automatic suggestions by confirming, revising, or rejecting them. In addition, the system minimizes errors by flagging unusual events. For instance, the trained specialist will check its classification for a significant deviation in an item's unit price. This approach ensures the data quality of the web-scraped data.

Expanding on the topic of machine learning techniques for product classification, Jahanshahi et al. (2021) compared text classification models for grocery product title classification in Turkey. They tested both traditional machine learning and advanced Natural Language Processing methods, such as BERT Devlin et al. (2018), ROBERTA Liu et al. (2019), and XLM Ma et al. (2020). They conclude that neural network-based models performed significantly better than traditional models, like SVM and XGBoost.

Considering the use of transformers architecture specific to the Portuguese language, which is the object of this research, Souza et al. (2020) released the BERTimbau model to the community after training the BERT model for Brazilian Portuguese. They related that the models improved the state-of-the-art in sentence textual similarity, recognizing textual entailment, and named entity recognition tasks, outperforming Multilingual BERT and confirming the effectiveness of large pre-trained Language Models for Portuguese. That said, BERTimbau is a good option for exploring this specific task and analyzing its performance compared with traditional models for the Portuguese language.

To utilize traditional machine learning models with text data, such as SVM and XGBoost, a crucial step involves converting words into numerical representations. One widely used approach is to generate word embeddings, compact vector representations designed to capture the semantic and syntactic relationships among words in a given language. These embeddings encode words' contextual meaning and associations, enabling the models to analyze and process textual information effectively. However, these models have predominantly been trained on English language data. In response to the need for Portuguese language models, Hartmann et al. (2017) trained 31 embedding models using FastText, GloVe, Wang2Vec, and Word2Vec on a large Portuguese corpus, including Brazilian and European variants. The study evaluated these models on syntactic and semantic analogies, POS tagging, and sentence semantic similarity tasks. Concerning the Semantic Similarity task, which best suits the product classification problem, the best results for European Portuguese were achieved by the Word2Vec CBOW model using 1,000 dimensions. It is important to emphasize that all the trained models mentioned in this study are readily accessible for download, facilitating their utilization in practical applications.

The product classification problem was also explored at the International Semantic Web Conference (ISWC) in 2020. The ISWC is a series of annual conferences focusing on developing and using the Semantic Web and related technologies, bringing together researchers, practitioners, and developers from academia, industry, and government to present and discuss the latest advances in the field and present challenges. One of the challenges proposed at the conference of 2020 was a product classification task to assign predefined product category labels to products sold on different websites. Zhang et al. (2020) presents the techniques utilized by the winners and reports that among the teams that submitted their system description paper, all competitors used Neural Network structures and the more recent transformer-based architectures or pre-trained language models. This suggests that Neural Network models can also be helpful in this type of task.

The product title consists of short texts. Therefore, it can be considered a short text classification problem, a crucial task in Natural Language Process. However, short texts have less contextual information than long texts, making them more ambiguous and complex for classification. In considering neural network architectures for text classification, the choice often tilts towards recurrent neural networks (RNNs) with long short-term memory (LSTM) for sequential data types like time series or text Hochreiter & Schmidhuber (1997). However, recent studies, such as those by Lee & Dernoncourt (2016) and Seo et al. (2020), have highlighted the comparative effectiveness of one-dimensional convolutional neural networks (CNNs) over LSTM-based RNN variants. These findings are particularly pronounced in short or unstructured texts, suggesting a potential advantage of CNNs in such scenarios.

Concerning an official classification for web-scraped data, Lehmann et al. (2020) proposed a classifier according to the ECOICOP schema on web-scraped products manually labelled from German retailers and transferred to the French language using cross-lingual word embeddings and Convolutional Neural Network. They compared its performance against a classifier trained on single languages and a classifier with both languages trained jointly. They demonstrated that zero-shot learning could be beneficial, especially when no manually labelled data is available in the early phases. In addition, they tested the quantity of data needed to build a single language classifier from scratch and explored multilingual training.

Martindale et al. (2020) presented a semi-supervised machine learning method for labelling and classifying web-scraped clothing data for CPI purposes. They proposed a semi-automated process to classify complex web-scraped data in consumer price statistics using a Label Propagation Algorithm. With the help of a minimal set of manually tagged data, the pipeline creates a large labelled dataset for training a standard classifier. In addition, the label propagation algorithm reduced the need for manual labelling and achieved a high accuracy rate in the test dataset.

Some statistical offices that use web-scraped data to calculate alternatives for CPI have already experimented with machine learning models for the classification task, but the methodology is private. However, many studies suggest that Neural Network-based models outperform traditional machine learning models. In addition, the studies reinforce the need to build a training set for the model, which can involve zero-shot learning using another language already labelled data or a semi-automated process for producing the classification for a small group of observations. Indeed, the classification process must constantly incorporate the evaluation of trained specialists in the field to improve the model results.

The literature on short text classification in Portuguese, particularly product classification, must be more extensive. This deficiency underscores the necessity of investigating this topic and assessing various models and scenarios for their effectiveness with Portuguese text. Moreover, implementing previously cited techniques such as a bilingual system, as suggested by Lehmann et al. (2020), would require a considerable number of instances (around 10,000) to achieve desirable performance. This fact highlights the significance of the initial step of gathering a set of labelled data. In addition, the semi-supervised learning method employed by Martindale et al. (2020) was targeted at clothing data. Given the diversity of food and beverage data surpasses that of clothing data, this approach seems less directly applicable to the current research. As such, the first stage in this research involves labeling an initial data set to train the models using traditional machine learning models such as SVM and XGBoost, transformer-based large language models, and neural networking models, focusing on exploring their performance with Portuguese product names. A detailed account of this methodology will be provided in the chapter to follow.

Chapter 3

Methodology

This research aims to classify items collected by web scraping of online supermarkets based on the ECOICOP schema, a hierarchical classification system described in United Nations Statistics Division (2018). This research will focus on the fifth-digit classification, as it is the approach for calculating inflation within the European Union. This level consists of 258 distinct categories, consistent across all euro area countries.

However, the scope of this research is limited to two broader categories relevant to products sold in online supermarkets: ”food and non-alcoholic beverages” and ”alcoholic beverages”. These two-digit categories encompass 71 of the 258 ECOICOP fifth-digit categories, accounting for 19 per cent of the inflation basket in the euro area, as mentioned in Lehmann et al. (2020). The categories and their official definition can be checked in Appendix A.

It is important to note that the distribution of products among these categories is highly imbalanced, which means that certain categories may have significantly more or fewer instances than others. This imbalance poses a challenge that needs to be addressed during the model development and evaluation process. The distribution of products per category can be check in the Appendix B.

An additional crucial challenge in this research was the absence of labelled data. Without a readily available training dataset, the task of providing the initial classification of products became paramount. This initial classification was the foundation for further training and refining the models. Indeed, the absence of labelled data posed a significant challenge in this research. To overcome this, the first step was to tackle a practical approach by generating a labelled dataset based on a single supermarket on a specific day, called supermarket ’A’, for confidentiality purposes. The methodology used to create the training dataset can be checked in section 3.3.

Once the labelled dataset for the first supermarket was created, it served as the training dataset for the models, such as XGBoost, SVM, LSTM, CNN, BERT, XLMRoberta, and BERTimbau. As the problem at hand involved short text classification, one of the key research questions of this project was to determine the optimal model for addressing this challenge.

To develop the XGBoost, and SVM models, Word2Vec word embeddings trained for the Portuguese language were employed, as referenced in Hartmann et al. (2017). These word embeddings represented words in a continuous vector space, allowing the models to capture semantic relationships and similarities between words in the Portuguese language. LSTM and CNN models were utilized with and without word embeddings to compare their results. By utilizing these various models and leveraging pre-trained word embeddings, the project aimed to compare their performance and identify the most suitable model for the short text classification task. The evaluation of these models was based on their accuracy and F1-score, enabling a comprehensive assessment of their effectiveness in classifying the products according to the defined categories. An overview of the modelling part can be checked in the section 3.6.

An iterative approach was defined to obtain the labelled data from all six supermarkets. This approach involved a step-by-step process to label the data and gradually improve classification accuracy. The process was summarized in the section 3.4.

The ultimate objective of this work was to develop a comprehensive pipeline capable of automatically classifying web-scraped food and beverage data into ECOICOP categories. This pipeline is critical in enabling these classified datasets for economic research in Portugal.

The following subsection of this chapter provides an overview of the dataset, ECOICOP categories, the process to obtain the training dataset, the iterative process to obtain the labels of the six supermarkets, the data preprocessing, the models used, and the evaluation metrics.

3.1 Data

The dataset used for the multi-level classification task was acquired from Banco de Portugal through daily web scraping of six online supermarkets in Portugal starting in December 2021. The names of the supermarkets used in this study will not be disclosed to ensure data confidentiality. Instead, they will be designated by letters such as A, B, C, D, E, and F when necessary.

Generally, the dataset comprises the following information about each product:

- ID: The internal identifier assigned by the supermarket.

- EAN-13: The European Article Number, a 13-digit numerical code that serves as a unique barcode identifier for the product.

- Name: The item's title on the website, which may include additional information such as brand, weight, or measurement unit.
- Brand: The name of the brand that produced the item.
- Category: The internal category assigned by the supermarket.
- Subcategories: Each supermarket has its own set of internal subcategories, which may range from 3 to 6.
- Price: This field contains information about the price, including price per quantity, price without discounts, and potentially other pricing formats. The format and details provided may vary across different supermarkets.
- Collection date: Indicates when the web scraping procedure was performed.

It is essential to consider the following points: the product ID is an internal identifier within the supermarket and is determined based on internal logistics. On the other hand, the European Article Number (EAN-13) code is a universal key used to identify products regardless of the supermarket. It provides a standardized identification method across different supermarket chains. Therefore, the same product from the same brand will have the same EAN-13 code, ensuring consistency in identification across various retailers. However, it is worth noting that for two supermarkets involved in this project, the EAN-13 code is unavailable on their respective websites.

Concerning the internal categories and subcategories fields, while some supermarkets can effectively identify the product type and aid in filtering food and beverage items from other items, it is only sometimes reliable and consistent. There are supermarkets where it may be impossible to determine the product's type solely based on internal categories, rendering this field useless. To address this challenge, it becomes crucial to develop a model to predict whether a product should be classified as food or beverage or categorized as another type, using only the product name and brand information. This enables accurate classification even when internal categories are not informative or available.

Another essential aspect to consider in not utilizing the internal category of the supermarket in the classification process is the need for more consistency and potential changes in these categories over time. Internal categories assigned by supermarkets may vary over time, leading to inconsistencies and inaccuracies in the classifications if these categories are used as model features.

In a first look at the dataset, it was possible to see that around 100 000 products are collected daily, meaning the whole dataset of one year contains more than 36 million products. Furthermore, a preliminary analysis of unique product names revealed that these products could not be easily grouped based on their names alone. Upon lowercasing the names of the products collected on December 21, it was discovered that fewer than 5 000 products have identical names.

3.2 ECOICOP

In the domain of data science, the quality of data holds significant importance. Within the scope of this study, a primary objective was to generate a high-quality training dataset to facilitate the development of an effective data classification workflow. Consequently, a substantial part of this research was dedicated to comprehending the European Classification of Individual Consumption according to Purpose (ECOICOP) classification system and establishing definitive guidelines for accurately classifying Portuguese products. To uphold the integrity of the data, a collaborative effort was undertaken with Banco de Portugal, whereby rigorous validation processes were implemented to ensure the quality of the data at each stage of the classification procedure.

Given the existence of 71 distinct categories and the inherent challenges associated with accurately classifying numerous products, meticulous efforts were undertaken to resolve any ambiguities and establish comprehensive definitions for each category. The adopted definitions were documented to ensure clarity and transparency throughout the project. Despite the time-consuming nature of this step, its significance cannot be overstated, as the overarching objective of this study is to utilize such data as a reliable and informative source for conducting economic research on the Portuguese economy.

In the ECOICOP classification scheme, there are a total of 61 classifications for food and 10 classifications for beverages. The specific categories within these classifications cover a range of products related to food and beverages. For a comprehensive list of all the categories, their descriptions, and the corresponding quantities of products classified within each category, please refer to Appendix A. This Appendix provides detailed information to supplement the understanding of the ECOICOP classification and the Appendix B provides the distribution of products across its various categories considering three different datasets used to train the models.

Since supermarkets typically offer a wide range of products, encompassing personal hygiene items, household cleaning supplies, kitchen utensils, gardening tools, toys, clothing, pet products, food supplements, and more, it was essential to identify the food and beverage items within the dataset. This step primarily involved filtering the main internal category assigned to the supermarket product, while in some instances, subcategories were also utilized as a filtering criterion. These products were classified as ”non-food and beverage products”. Additionally, food products categorized as a combination of multiple items (i.e., a basket of different products) and food supplements were also classified as ”non-food and beverage products” due to the absence of a specific ECOICOP classification for such cases.

Therefore, the problem was defined as a multi-classification task with 72 categories. Among these categories, 71 correspond to the specific ECOICOP categories that are the focus of this work, and the remaining category was designed to capture all the items collected that do not fall into any of the 71 ECOICOP categories.

3.3 Training Dataset

A deep understanding of the ECOICOP classification was needed to create the training dataset, and the official documentation of the 71 categories within the ECOICOP classification served as the guideline for this classification process.

The first classification was based on defined rules, primarily using the first word of each product. For instance, if the first word of a product is ”spaghetti,” it would be classified under the category ”Pasta products and couscous” according to the ECOICOP classification. Similar rules were established to estimate the category based on the first word of each product. The supermarket selection for the initial classification was based on the supermarket with the most accurate internal category structure aligned with ECOICOP. The presence of an internal category field within the dataset played a significant role in facilitating and accelerating the classification process. This field provided valuable insights into understanding the items, and in some cases, the internal categories corresponded directly to those used in the ECOICOP schema. It is essential to highlight that this first supermarket contained around 27 thousand products.

Addressing the categorization of products within each unique specific category was a complex task. With some categories overlapping, careful consideration and decision-making were necessary to determine which products should be included in each category. Throughout this process, decision criteria were established and thoroughly documented. This documentation serves multiple purposes, including facilitating a comprehensive understanding of the project’s classification methodology and providing a reference guide for potential future improvements or refinements. This material ensures transparency, consistency, and reproducibility of the project. This material was produced along Banco de Portugal for internal consultation and is available only in Portuguese.

The labelled dataset created for this project underwent a rigorous review process conducted by Banco de Portugal. Also, the classification criteria and decisions were subject to thorough discussions and consultations at each project stage. The project benefited from their expertise and domain knowledge in the field by involving Banco de Portugal in the review process. Their input helped ensure the accuracy and reliability of the labelled dataset and the classification criteria employed.

3.4 Iterative Process

To classify the products from other supermarkets, an iterative process was employed. Initially, the best-performing model, determined from the evaluation in the manually labelled dataset of supermarket A, was utilized to classify the products from the second supermarket, defined as supermarket B. The model's predictions were subject to further review, and any inconsistencies were manually corrected. After the manual review and corrections, all the products from the second supermarket were appropriately labelled. The labelled dataset from the second supermarket was combined with the labelled dataset from the first supermarket to retrain the best model aiming to enhance its performance.

The retrained model was subsequently utilized to predict the categories of products from the third supermarket. The classifications made by the model were further reviewed by specialists from Banco de Portugal, who manually corrected any inconsistent classifications. This iterative process of training, reviewing, and manually correcting inconsistencies was repeated until all six supermarkets had the correct ECOICOP category labels assigned to their products. This meticulous approach ensured the accuracy and reliability of the final classifications across all supermarkets. By incorporating expert review and continuous improvement through iteration, the project achieved robust and consistent ECOICOP category labelling for the products across the entire dataset, ensuring the data quality of the input to machine learning models. This process is presented in the figure below.

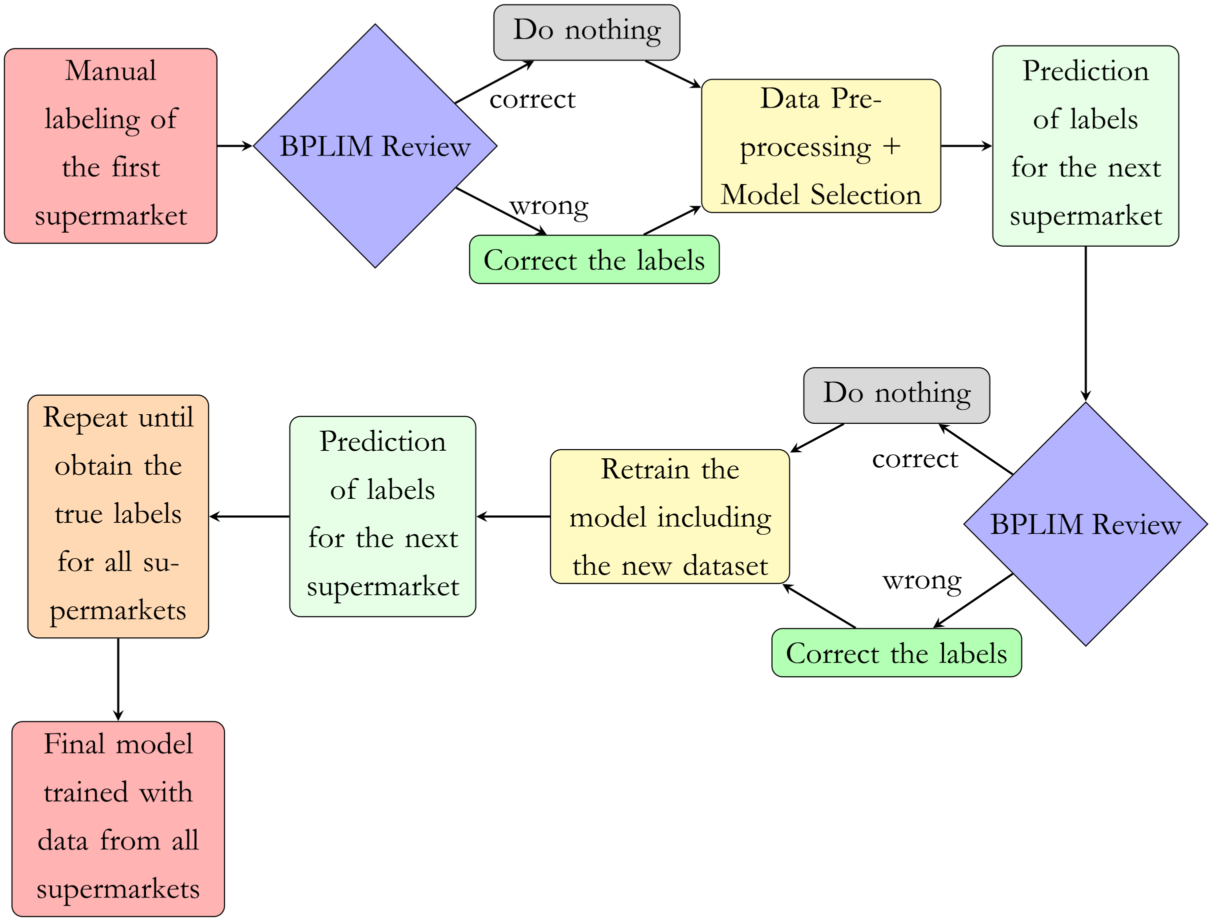

Manual labeling of the first supermarket
BPLIM Review
Do nothing
correct
wrong
Correct the labels
Data Pre-processing + Model Selection
Prediction of labels for the next supermarket
BPLIM Review
Do nothing
correct
wrong
Correct the labels
Retrain the model including the new dataset
Prediction of labels for the next supermarket
Repeat until obtain the true labels for all supermarkets
Final model trained with data from all supermarkets

Figure 3.1: Classification Process

The final model was retrained after accumulating a substantial dataset of almost 100 000 labelled products from the six supermarkets. The next step involved labelling all the products collected throughout 2022, totalling more than 36 million items. In this labelling process, an initial analysis was conducted to identify products with an EAN-13 code that matched an already classified product. For these cases, the ECOICOP classification was assumed based on previous labelling, and there was no need for the model's prediction.

However, the trained model was applied to classify products that had not appeared in the dataset. This enabled the model to classify a large volume of new products. By utilizing this two-step approach, the project ensured that products with matching EAN-13 codes were classified based on previous knowledge. In contrast, newly encountered products were classified using the trained model's predictions and then reviewed by BPLIM.

3.5 Data Preprocessing

The preprocessing step was critical in standardizing the product names for input into the models. Each supermarket had its unique format for presenting item names, which included variations such as including the brand within the name, specifying the unit (e.g., ”ml,” ”kg,” ”l”), or indicating the minimum quantity for sale (e.g., ”- Unidade 0.4 Kg”). To address these variations and ensure consistency, several preprocessing techniques were applied. Firstly, all product names and brands were transformed to lowercase to eliminate case sensitivity. Next, a regular expression function removed units, numerical values, and special characters from the product names. This step aimed to eliminate unnecessary noise and focus on the essential textual information.

However, there was an exception regarding the removal of numbers. Non-alcoholic beers, for example, sometimes included only ”0%” or ”0.0%” in their names to indicate their category as ”Low and non-alcoholic beer”. Therefore, when the product contains the word ”cerveja”, which means beer in Portuguese, numbers were not removed, as they held significance for classifying the product correctly.

Additionally, the preprocessing step included detecting if the brand name was already present in the product name. If the brand was not found, it was appended to the name. Incorporating the brand information into the product name proved valuable for the classification process, providing additional context for the models. Through these preprocessing techniques, the product names were standardized and prepared as input for the models, ensuring a consistent and coherent data representation.

The following diagram illustrates the preprocessing pipeline culminating in the tokenization step, which converts the text data into individual units and serves as the input for the models presented in the following subsection.

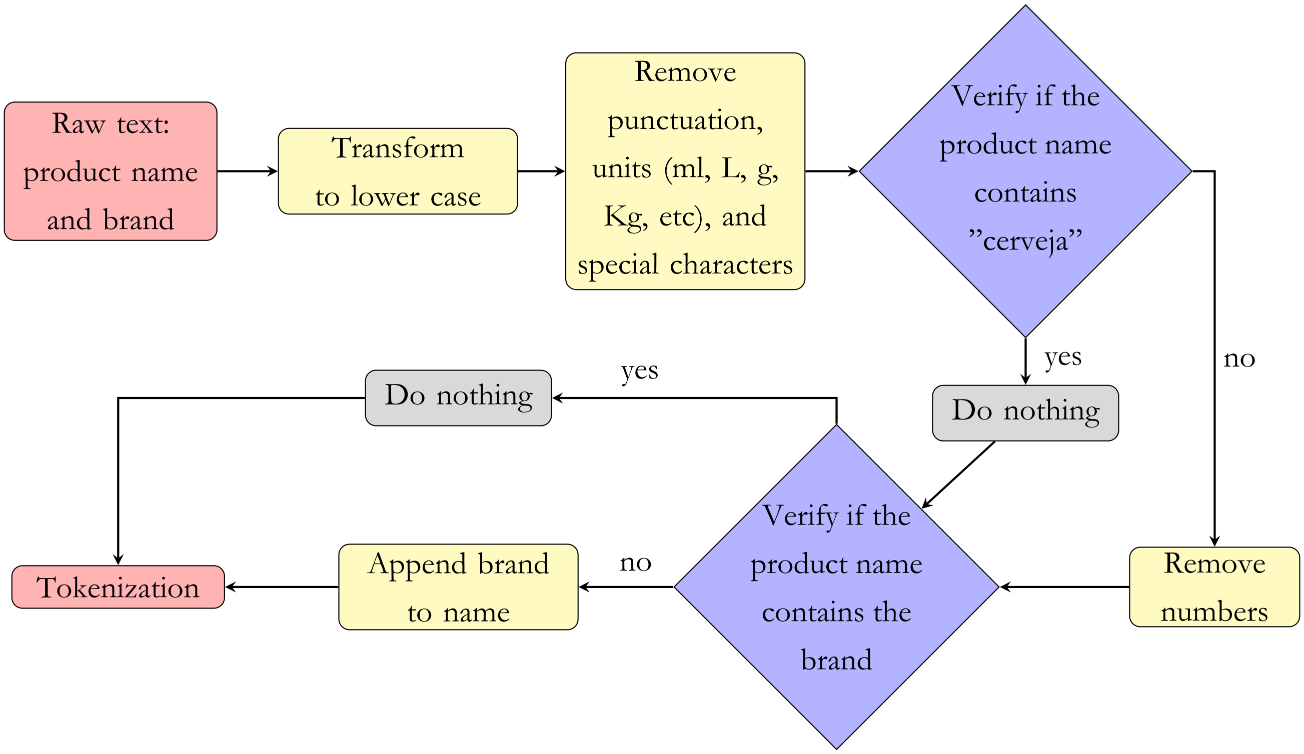

Raw text: product name and brand
Transform to lower case
Remove punctuation, units (ml, L, g, Kg, etc), and special characters
Verify if the product name contains ”cerveja”
yes
no
Do nothing
Remove numbers
Verify if the product name contains the brand
yes
Do nothing
no
Append brand to name
Tokenization

Figure 3.2: Preprocessing Pipeline

3.6 Models

To train the models, the dataset was divided with an 80-20 split for training and testing, respectively. This split was stratified based on the category to ensure a comprehensive representation of the products. To guarantee the correct distribution in train and test datasets, products of categories with less than five items were augmented based on the existent names. This approach was selected due to the highly imbalanced nature of the dataset. Imbalanced datasets have significantly unequal class distributions, which can lead to biased model performance and inaccurate classification results. For instance, considering the first labeled dataset of supermarket A, ”Wine from grapes” contains more than 4100 observations, and the category ”Beer-based drinks”, which includes products like beer with soda (”panache”), and beer with Coca-Cola, contains only 1 observation. Using a stratified sampling technique, the train and the test datasets contain the same class distribution. This ensures that each fold represents a proportional representation of the different classes, mitigating the potential bias caused by class imbalance.

Concerning model input, it is essential to note that both XGBoost and SVM models require numerical data instead of string inputs for data processing. To this end, this project utilized a Word2Vec word embedding model to convert textual data into numerical representations that these models can comprehend. The Word2Vec model employed in this project was tailored explicitly for the Portuguese language, leveraging the Continuous Bag-of-Words (CBOW) method as outlined by Hartmann et al. (2017). In the CBOW technique, the model predicts a missing word from a sequence, thereby learning significant word representations from their context. The adopted Word2Vec model produced word embeddings with a dimension size of 1000, indicating that each word was represented as a dense vector comprising 1000 dimensions. These embeddings encapsulated semantic relationships and contextual nuances, thus enabling the models to discern and interpret the inherent textual patterns.

In addition, the Portuguese word embeddings were also integrated into both Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures, establishing a basis for comparative evaluation of these models' performance with and without the usage of pre-trained embeddings. LSTM and CNN models can use pre-trained embeddings as input, providing a preliminary semantic understanding of the data. Alternatively, these models can generate their embeddings by initializing with random values and refining them during learning. The comparative analysis thus aimed to explore the influence of these different methodologies on the performance of these models in short text classification tasks.

On the other hand, large language models such as BERT, XLM-RoBERTa, and BERTimbau have tokenizer methods that transform words into vectors. These models utilize tokenization techniques specifically designed for their architectures, which means that the tokenization method of these models can split the text into smaller units, such as words, and then represent it as numerical vectors that capture the semantic and contextual information of the tokens, allowing the models to process and understand the text. Being transformers-based, these models exhibit a critical architectural feature: self-attention mechanisms Vaswani et al. (2017), which allow them to capture contextual relationships between words in a sentence, regardless of their positions. The tokenization methods produce tokenized inputs fed into the transformer-based models, where the self-attention mechanism generates embeddings based on the context. These models were fine-tuned for the classification task by adding a final layer to the pre-trained model with the ECOICOP labels.

Below is a brief description of each model used in this study:

- XGBoost (Extreme Gradient Boosting): XGBoost is a powerful gradient boosting framework known for its high performance and efficiency. It uses a set of decision trees and employs a gradient-boosting algorithm to optimize model performance. XGBoost's ability to handle sparse, high-dimensional data makes it particularly suitable for short text classification, making it suitable for text-based tasks involving many features. For instance, Qi (2020) compared classical machine learning models for classifying theft crime data of a city based on the brief description of the theft crime, which they defined as a short text classification task, and concluded that the XGBoost algorithm performed better than KNN, Naïve Bayes, SVM, and Gradient Boosting Decision Tree algorithms for this task.

- SVM (Support Vector Machines): SVM is a widely used supervised learning algorithm that aims to find an optimal hyperplane in a high-dimensional space to distinguish different classes. It is adequate for short text classification tasks because it can handle non-linear data and is robust against overfitting. SVM works well with limited training data and is known for its ability to handle large feature spaces. García-Méndez et al. (2020) highlights the high accuracy of SVM in comparison with other models considering complexity and computing time for short text classification on a specialized labeled corpus for identifying banking transaction descriptions. Also, Adhi et al. (2019) provided a systematic literature review of short text classification on Twitter and concluded that SVM is the most widely used method for this task.

- LSTM (Long Short-Term Memory): LSTM is a Recurrent Neural Network (RNN) designed to capture long-term dependencies in serial data. This is relevant for text classification, as it can effectively model the sequential nature of a text by storing important information in longer sequences. LSTM has shown promising results in various natural language processing tasks, including text classification. For instance, Xiao et al. (2018) proposed a classifier based on Word2Vec and LSTM for patent text classification that presented impressive results, with a classification accuracy rate of 93.48%, overcoming K Nearest Neighbor and Convolutional Neural Network models.

- CNN (Convolutional Neural Network): CNNs are deep learning models exhibiting remarkable capabilities in short text classification tasks. CNNs employ convolutional layers to discern important local contextual features within a given text, further summarized by pooling layers into fixed-length vectors. This makes them adept at managing different text lengths and positions, particularly useful for short text classification. J. Wang et al. (2017) used CNN of seven layers and word embeddings on five different datasets, such as TREC, Twitter, AG news, Bing and Movie Review, and states that their model outperforms the state-of-the-art methods for short text classification.

- BERT (Bidirectional Encoder Representations from Transformers): BERT is a state-of-the-art pre-trained language model. It utilizes a transformer architecture and is trained on a large corpus of unlabeled text data. BERT's relevance for text classification lies in its ability to capture contextual information and semantic relationships between words. By leveraging its pre-trained representations, BERT can learn complex text patterns and provide accurate classification results. Zahera & Sherif (2020) described their submission to the semantic web challenge on mining the product data collected from various websites to predict its category. They presented a fine-tuned BERT model for this task and concluded that the model is an excellent baseline to benchmark the task of automatic product classification.

- XLM-RoBERTa (Cross-lingual Language Model - RoBERTa): XLM-RoBERTa is another transformer-based language model variant extending BERT's capabilities to multilingual text. It is trained on a vast amount of multilingual data and can handle short text classification tasks in various languages. XLM-RoBERTa's relevance lies in its ability to leverage pre-trained representations across different languages, allowing it to generalize well and perform highly on text classification tasks involving multilingual data. This model was also used for classifying products to the semantic web challenge Kertkeidkachorn & Ichise (2020), the team states that the Roberta-large model had the best precision among all the models.

- BERTimbau: BERTimbau is a BERT-based language model designed explicitly for Portuguese. It is pre-trained on a large corpus of Brazilian Portuguese text, enabling it to capture the language's specific linguistic nuances and patterns. BERTimbau claims to achieve state-of-the-art performances on three downstream NLP tasks for Brazilian Portuguese: Named Entity Recognition, Sentence Textual Similarity, and Recognizing Textual Entailment. Numerous studies, including Souza & Souza Filho (2022) and Santana et al. (n.d.), demonstrate the exceptional performance of BERTimbau in text classification tasks for the Portuguese language. They found that a fine-tuned BERTimbau model could accurately distinguish textual differences across various categories and outperformed all other techniques evaluated in these studies.

It is essential to highlight that large language models such as BERT, XLM-RoBERTa, and BERTimbau, were pre-trained on massive amounts of unlabeled text data and are available through Hugging Face, an open-source software library and community that focuses on democratizing and accelerating the adoption of NLP models trough the ”transformers” library Wolf et al. (2020). BERT and XLM-RoBERTa were trained in more than 100 languages, making them versatile models for multilingual NLP tasks. On the other hand, BERTimbau was explicitly designed for the Portuguese language and was pre-trained on a significant corpus of Portuguese text using techniques similar to BERT and XLM-RoBERTa.

3.7 Evaluation Metrics

The evaluation of the classification models involved the following metrics:

- Accuracy: Accuracy measures the proportion of correctly classified samples over the total number of samples. It provides an overall measure of the model's correctness in predicting the correct categories.

- Macro F1 Average: F1 score is a metric that combines precision and recall. Macro F1 average calculates the F1 score for each class independently and then takes the average across all classes. It provides a balanced assessment of the model's performance across all categories, considering precision and recall, irrespective of their representation in the dataset.

These metrics were chosen to evaluate the model's performance comprehensively, considering overall accuracy and the ability to classify each item category correctly. Accuracy provides a general measure of correctness, while the macro F1 average considers the model's performance across all categories, giving equal weight to each category. In addition to accuracy and macro F1 average, the processing time to train the model was also considered.

By considering accuracy, macro F1 average, and processing time as evaluation metrics, the project aimed to assess the model's overall performance, ability to classify individual categories, and data processing efficiency. These metrics provided a comprehensive view of the model's strengths and limitations, helping to guide improvements and optimize the classification process.

3.8 Software

Jupyter Lab was used as the integrated development environment (IDE) for this research, with Python as the chosen programming language. Multiple packages were employed for various purposes. The *transformers* package was utilized to obtain pre-trained BERT-based models, and

Keras included in *TensorFlow* package was used to construct the CNN, LSTM and BERT-based models. The *gensim* package was used for loading and utilizing pre-trained Word2Vec models to capture semantic relationships between words. Other packages, such as *pandas*, facilitated efficient data frame manipulation. The *scikit-learn* package was instrumental in performing label encoding, train-test splitting, and model evaluation and provided the SVM model. Also, the *XGBoost* package was utilized for implementing and evaluating the XGBoost model. The *re* package also facilitated preprocessing steps involving regular expressions. This combination of packages enabled data handling, model development, and evaluation throughout the research workflow.

3.9 Assumptions and Limitations

The assessment of the models, which included preprocessing time, was conducted after executing all experiments on a system equipped with an Intel(R) Core(TM) i7-10750H CPU operating at 2.60GHz (with a base speed of 2.59 GHz) and featuring 16 GB of RAM. Concerning this topic, the large language models utilized (BERT, XLM-Roberta, and BERTimbau) are the ”base” versions, which are more computationally efficient. Nevertheless, achieving enhanced results is possible by opting for these models' more resource-intensive ”large” versions.

Hyperparameter optimization was undertaken for the top-performing models using grid search, but it did not notably enhance their performance. An interesting exception to this was the BERT-base models. They required a learning rate not exceeding 2×10^{-5} to demonstrate impressive results, as higher learning rates led to worse outcomes. This underscores the critical role of selecting an appropriate learning rate for specific architectures like BERT in optimizing model performance.

Given the large number of categories, this study will not individually assess and analyze the performance of each category. Instead, the focus will be on the broader picture, examining overall performance across all categories. The next chapter presents the research questions guiding this research and the results.

Chapter 4

Results and Discussion

4.1 Research Questions

As said before, the primary objective of this research was to automate the classification of supermarket products collected by web scraping according to the European Classification of Individual Consumption According to Purpose (ECOICOP) standards. A central research question guiding this study was identifying the most effective model for this classification task. To answer this question, an extensive evaluation of various machine learning and deep learning models was conducted. These models encompassed traditional machine learning models such as Support Vector Machines (SVM) and XGBoost, as well as more advanced deep learning models, such as Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and large language models like BERT, XLM-RoBERTa, and BERTimbau.

As part of this evaluation, pre-trained word embeddings for the Portuguese language were incorporated to ascertain whether this strategy could enhance the performance of the deep learning models. This formed the second research question: Does the application of word embeddings improve the performance of CNN and LSTM in this specific task?

The third research question was centred around the use of large language models. Specifically, the question was whether models trained in a specific language, such as BERTimbau for Portuguese, outperform cross-language models like BERT multilingual base model, a pre-trained model on the top 104 languages, and XLM-RoBERTa, a model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

Finally, a key question was examined: ”How effectively can the models trained on data from one supermarket generalize when applied to fresh, unseen data from another supermarket?” To answer this, the top-performing models, based on a 94% accuracy threshold, were reassessed using another supermarket dataset.

4.2 Results

The subsequent table shows the models' performance metrics, utilizing a dataset gathered from a singular supermarket on a specific day in the year 2021. This data was employed for the preliminary assessment of these models. The dataset was strategically partitioned with 80% allocated for model training and the remaining 20% designated for testing. The performance evaluation of the models was executed based on key metrics such as accuracy, Macro F1 score, and training processing time.

Model	**Accuracy**	**F1 Macro**	**Time Processing (min)**
SVM + Word Embeddings	90.5%	74.1%	5.7
XGBoost + Word Embeddings	86.2%	65.4%	23.3
LSTM	94.1%	65.9%	2.5
LSTM + Word Embeddings	95.3%	77.3%	7.0
CNN	96.8%	**87.3**%	2.1
CNN + Word Embeddings	95.4%	86.5%	**1.3**
BERTimbau	**97.3**%	72.2%	192.5
BERT	94.9%	63.9%	223.0
XLM-RoBERTa	92.2%	58.9%	241.6

Table 4.1: Models' performance comparison on the same supermarket

Based on these findings, the research questions can be addressed as follows:

Which is the best model for the short text classification task?

- The traditional machine learning models had the worst accuracy among all models, despite that, SVM enhanced with word embeddings achieved an accuracy of over 90% (90.5%) and an F1 Macro score of 74.1%. In contrast, XGBoost with word embeddings manifested slightly diminished efficacy, acquiring an accuracy of 86.2%, and an F1 Macro score of 65.4%. It's worthy of note that SVM demonstrated superior time efficiency, achieving its results in a mere 5.7 minutes compared to the 23.3 minutes required by XGBoost.

- Concerning deep learning models, LSTM and CNN displayed outstanding performances, with and without word embeddings. Specifically, the CNN model presented a robust accuracy of 96.8% and an impressive F1 Macro score of 87.3%, obtained in only 2.1 minutes. Remarkably, word embeddings only slightly affected CNN's accuracy, reducing it to 95.4%, but the processing time was cut further to 1.3 minutes.
- As for the Large Language Models, the Portuguese-specific BERTimbau surpassed others in accuracy, reaching 97.3%, the best result among all models tested. Despite this, its F1 Macro score of 72.2% lagged behind the CNN models. Moreover, the processing time for BERTimbau was significantly longer, clocking in at 192.5 minutes.

What is the effect of pre-trained word embeddings on CNN and LSTM performance?

The findings suggest that using pre-trained word embeddings notably enhanced the performance of the LSTM model, with the accuracy increasing from 94.1% to 95.3% and the F1 Macro score rising from 65.9% to 77.3%. On the other hand, the impact on the CNN model was less noticeable. When word embeddings were applied, the accuracy of CNN showed a marginal decrease from 96.8% to 95.4%, and there was a slight reduction in the F1 Macro score. Regarding processing time, the introduction of word embeddings more than doubled the processing duration for LSTM while reducing the processing time of CNN by 40%

Do large language models specifically trained for a particular language demonstrate superior performance to cross-language models?

Regarding 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). BERTimbau also proved superior in processing time, taking the least time among this category at 192.5 minutes. In comparison, BERT required 223.0 minutes, and XLM-RoBERTa took 241.6 minutes.

What is the generalizability of these models when applied to new data from another supermarket?

Upon completing the initial analysis and evaluating the models using a split dataset from one supermarket, a new research question emerged: How would these models perform when applied to a new dataset from a different supermarket? To explore this question, models that had achieved an accuracy threshold of 94% in the previous evaluation that were trained exclusively on the dataset from supermarket A were then tested on data from supermarket B. This step was crucial to ensure that the high accuracy in the first evaluation was not due to overfitting. Therefore, the CNN, LSTM (both with and without word embeddings), BERT, and BERTimbau models underwent further evaluation.

Model	**Accuracy**	**F1 Macro**
LSTM	88.3%	55.1%
LSTM + Word Embeddings	90.2%	63.8%
CNN	91.3%	**74.6**%
CNN + Word Embeddings	88.8%	71.0%
BERTimbau	**92.1**%	63.9%
BERT	90.2%	58.2%

Table 4.2: Models' performance on new data

Several insights can be derived upon comparing the models' performance in the new data. First, every model experienced a dip in performance when transitioning from the training data to the unseen data, as expected. This observation highlights the challenges inherent to machine learning models in generalizing to unseen data.

Despite being the best performers on the training dataset, CNN models witnessed the most significant drop in accuracy, from 96.8% to 91.3%, and F1 Macro score, from 87.3% to 74.6%. Although this drop is significant, the CNN model remains the top performer on the new dataset, underscoring its robustness despite the unseen data. In contrast, the LSTM models, with and without word embeddings, demonstrated a more moderate decline in performance metrics. However, they started from lower baseline performance, and the drop brought them to less competitive performance levels in the new dataset. The large language models, BERT and BERTimbau, also saw decreases in their performance metrics, indicating some challenges in generalization. The F1 Macro score, particularly sensitive to class imbalances in the data, showed more pronounced drops for these models, suggesting a potential weakness in handling new data with different class distributions.

In conclusion, the Convolutional Neural Networks (CNN) model emerged as the top performer for the product classification task. Upon acquiring the true labels from all six supermarkets, the model was retrained and demonstrated superior performance, achieving an accuracy of

96.60% and an F1 Macro score of 92.19%. These outcomes underscore the crucial role of data diversity and the need for a sufficient volume of observations within each category to ensure robust and accurate model performance.

4.3 Discussion

Considering the results, highlighting a few key points is essential. Firstly, deep learning and large language models notably outperformed traditional machine learning models such as SVM and XGBoost in the context of short text classification for supermarket products. Even though SVM had a higher F1 Macro score than large language models, the latter showed superior accuracy. Concerning the most appropriate model, the CNN model exhibited impressive performance in short text classification, achieving high accuracy and F1 Macro scores. Additionally, the CNN model demonstrated notable advantages in processing time, taking only 2 minutes to be trained in a dataset with approximately 27 thousand products. In comparison, BERTimbau, although delivering strong performance, required over 3 hours of training in the same dataset. This significant difference in processing time further reinforces the efficiency of the CNN model, making it a more time-effective choice for short text classification tasks in scenarios where quick model training and deployment are crucial or computational resources are limited. Therefore, considering its strong performance and efficient processing time, the CNN model emerged as a favourable choice for short-text classification in this study.

Secondly, the performance of LSTM presented substantial improvements with the incorporation of pre-trained word embeddings, especially regarding the F1 Macro score, enabling the model to capture semantic meaning and context more effectively. This shows the potential benefit of leveraging pre-existing knowledge from word embeddings for text classification problems, especially using LSTM models. However, it is essential to note that using pre-trained word embeddings significantly improved the performance of LSTM models. It did not yield the same impact on CNN models. This suggests that the effectiveness of pre-trained word embeddings may vary depending on the specific task and model architecture. Therefore, carefully evaluating pre-trained word embeddings in each task is crucial to determine their potential benefits and impact on model performance.

Thirdly, the results demonstrate the significance of language-specific large language models. In particular, the model trained specifically for the Portuguese language, BERTimbau, outperformed cross-language models such as BERT and XLM-RoBERTa. Notably, 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.

Lastly, while the results of this study provide valuable insights for the task of supermarket product classification, it is necessary to acknowledge that these findings might not be generalized to other problems. The performance of different models can vary greatly depending on the specifics of the dataset and the problem at hand. Therefore, evaluating multiple models in the context of a specific task is always recommended.

Chapter 5

Conclusion

The primary aim of this Master's dissertation was to lend robust technological support to Banco de Portugal Microdata Research Laboratory - BPLIM, aligning them with the evolving paradigms of data management and analytics. As the world progresses toward digitalization, the volume of data generated has exponentially increased in recent years, and new data sources are emerging. This study effectively recognized the need for BPLIM to tackle this new data landscape and proposed solutions drawn from the latest developments in data analytics.

The first stage of the research revolved around a thorough literature review that concentrated on big data analytics. This literature review was essential to BPLIM and served as an informative guide, offering a deep understanding of the various aspects associated with managing the growing influx of structured and non-structured data. This research has given BPLIM the know-how to make sense of different database systems and their use cases. It also gives a simple rundown of file formats and the main frameworks for processing data in a cluster environment.

The evolution in data volume, velocity, and variety necessitates a parallel advancement in the methodologies employed to manage, process, and analyze it. The literature review has equipped BPLIM with a solid understanding of the necessary tools and techniques for leveraging big data analytics, thus ensuring that they are well-prepared to face the emerging paradigms in the data ecosystem.

The second phase of this research focused on exploring the most effective machine learning models to automate a multi-classification task of products acquired through web scraping from online supermarkets. BPLIM has been collecting this data daily for over one and a half years. However, due to the absence of an official classification for each collected product, this data had yet to be utilized for internal studies. Therefore, the practical issue addressed in this research was to devise an efficient method to automate this classification task, conforming to the 71 European

Classification of Individual Consumption categories according to Purpose - ECOICOP.

The aim was to enable this newly sourced data to contribute to producing official statistics about the Portuguese economy. The benefits of web scraping in generating official statistics are well established, showing several advantages over traditional data collection techniques, such as high frequency, timeliness, and granularity. The most prominent use case emerges while handling the Consumer Price Index, an aspect that numerous statistical institutes and central banks have already explored.

However, a substantial impediment in this project was the prerequisite for labelled data that could be utilized to train the machine learning models. Consequently, the initial step involved labelling data from a specific day from one supermarket, intended to serve as the training dataset for predicting data from other supermarkets on the same day. This labelling process necessitated the expertise of a BPLIM specialist who meticulously reviewed the labels. Although this task was time-intensive, it was an essential part of accelerating the classification process for other supermarkets. Upon acquiring the labelled dataset from the initial supermarket, it was divided into training and testing subsets, and various models were subsequently evaluated. The most effective model derived from this evaluation was employed to automate the labelling of other supermarkets. These newly assigned labels underwent a rigorous review by BPLIM, which permitted an assessment of the quality of the model's predictions. This step ensured the adequate performance of the model on new data and guaranteed the reliability of the derived classifications.

Convolutional Neural Networks (CNN) and BERTimbau emerged as highly accurate, achieving over 91 % accuracy among the models evaluated on the new data. However, CNN displayed a higher F1 macro score of 74%, indicating that this model could more efficiently generalize to categories with fewer data instances. A crucial aspect considered during this research was the processing time of the models. While large language models required over three hours to deliver a model on a training dataset of 27 thousand instances, CNN only required two minutes.

Ultimately, the Convolutional Neural Network (CNN) model was retrained with all data effectively labelled, utilizing nearly 100,000 instances. Impressively, the model demonstrated high efficacy, attaining an accuracy of 96.60 % and an F1 Macro score of 92.19 %. Following the successful retraining, the model was then deployed to classify all products collected throughout 2022, which comprised over 200 thousand unique products. The model's robust performance and ability to efficiently process large volumes of data make it an excellent tool for classifying vast quantities of daily collected products. By employing this model, BPLIM can streamline the data classification process and use this new data source to produce official statistics. Appendix C contains a letter from BPLIM, outlining their perspective on the importance of this research for their projects. The letter also provides valuable feedback on the internship that served as the foundational basis for this research. This document highlights the tangible real-world impact of this study, affirming its relevance and value for Banco de Portugal.

The originality of this study lies in its pioneering methodology for classifying products gathered through web scraping based on the ECOICOP official categorization. Despite the handful of published studies in this domain, none have specifically focused on the Portuguese language, highlighting the uniqueness of this work. The advent of this research opens up new pathways for Portuguese-language studies in product classification for official statistics, fostering significant opportunities for the research community. In addition, this study critically evaluates various models for short text classification within the context of the Portuguese language, an area that has remained unexplored in the literature. Therefore, this study addresses a crucial knowledge gap and establishes a benchmark for future explorations, thereby contributing a valuable foundation for ensuing advancements in this research area.

5.1 Future Work

The evolution of this research presents a multitude of promising paths for further exploration and enhancement. An instinctive development of this work would be the expansion of the classification spectrum beyond just food and beverage products to incorporate all 258 categories within the European Classification of Individual Consumption according to Purpose (ECOICOP). Significantly, nearly half of the unique products acquired through the web scraping routine from the supermarkets do not fall under the categories of food or beverages. This unclassified segment represents an untapped repository of valuable information that could further enrich BPLIM's research scope.

Considering the remarkable performance of the Convolutional Neural Network (CNN) model in classifying food and beverage items, it is logical to foresee similar success when the model is applied to other product categories. Enlarging the model's scope to classify items across the entire ECOICOP ambit could offer a more comprehensive and granular perspective of consumption behaviours. Also, it would be interesting to explore the potential of hybrid models such as BERT-CNN, BiLSTM, and CNN-BiLSTM-Attention, which have shown promising results in text classification tasks.

Concerning the necessity for labelled data to initiate the training process, one possible approach that holds potential is the employment of prompt engineering in conjunction with large language models, such as GPT-3. This approach could provide an initial automated classification, thus expediting the initial stages of the labelling process. Naturally, any classification obtained through this method should undergo a rigorous review by BPLIM specialists to ensure the highest data integrity and quality.

Lastly, an area for future focus is continuously updating and improving the model's accuracy as additional labelled data becomes available. Considering the fluid nature of data and consumption trends, periodic retraining of the model is necessary to uphold its high performance and relevance. In conclusion, the potential future work from this research holds significant promise for further enhancing BPLIM's data analysis capabilities, allowing for deeper insights across a broader range of consumption categories. By leveraging these opportunities, BPLIM can stay at the cutting edge of data analytics, providing crucial insights that contribute to shaping economic policies and strategies.

Bibliography

Adhi, B., Saskiah, D., & Widodo, W. (2019, 03). A systematic literature review of short text classification on twitter. *KnE Social Sciences*, *3*, 625. doi: 10.18502/kss.v3i12.4134

Aluko, V., & Sakr, S. (2019, 12). Big sql systems: an experimental evaluation. *Cluster Computing*, *22*, 1347-1377. doi: 10.1007/s10586-019-02914-4

Aparicio, D., & Bertolotto, M. I. (2020). Forecasting inflation with online prices. *International Journal of Forecasting*, *36*(2), 232-247. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207019301530 doi: https://doi.org/10.1016/j.ijforecast.2019.04.018

Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., & Trunfio, P. (2022, 12). Programming big data analysis: principles and solutions. *Journal of Big Data*, *9*. doi: 10.1186/s40537-021-00555-2

Bertolotto, M. (2016, 09). The perils of using aggregate data in real exchange rate estimations. *Social Science Research Network*. doi: https://ssrn.com/abstract=2882339orhttp://dx.doi.org/10.2139/ssrn.2882339

Bluhm, B., & Cutura, J. A. (2022). Econometrics at scale: Spark up big data in economics. *Journal of Data Science*, 413-436. doi: 10.6339/22-jds1035

Cavallo, A. (2017, 01). Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review*, *107*, 283-303. doi: 10.1257/aer.20160542

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research [Article]. *Journal of Economic Perspectives*, *30*(2), 151 – 178. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84976505306&doi=10.1257%2fjep.30.2.151&partnerID=40&md5=00c141df467ac2685259c24e3f9954e0 (Cited by: 95; All Open Access, Bronze Open Access, Green Open Access) doi: 10.1257/jep.30.2.151

Dean, J., & Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th conference on symposium on operating systems design implementation - volume 6* (p. 10). USA: USENIX Association.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv. Retrieved from https://arxiv.org/abs/1810.04805 doi: 10.48550/ARXIV.1810.04805

Eurostat. (2020a). European statistical system handbook for quality and metadata reports [Computer software manual]. Retrieved from https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783?t=1583397712000 doi: 10.2785/666412

Eurostat. (2020b). Practical guidelines on web scraping for the hicp [Computer software manual]. Retrieved from https://ec.europa.eu/eurostat/documents/272892/12032198/Guidelines-web-scraping-HICP-11-2020.pdf

Faridoon, A., & Imran, M. (2021, Nov.). Big data storage tools using nosql databases and their applications in various domains: A systematic review. *COMPUTING AND INFORMATICS*, *40*(3), 489–521. Retrieved from https://www.cai.sk/ojs/index.php/cai/article/view/2021_3_489 doi: 10.31577/cai_2021_3_489

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. Retrieved from https://www.sciencedirect.com/science/article/pii/S0268401214001066 doi: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

García-Méndez, S., Fernández-Gavilanes, M., Juncal-Martínez, J., González-Castaño, F. J., & Seara, □ PBI B. (2020). Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus. *IEEE Access*, *8*, 61642-61655. doi: 10.1109/ACCESS.2020.2983584

Gohil, A., Shroff, A., Garg, A., & Kumar, S. (2022). A compendious research on big data file formats. In (p. 905-913). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICICCS53718.2022.9788141

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 Fore-*

casting, *34*(2), 225-234. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207018300013 doi: https://doi.org/10.1016/j.ijforecast.2017.12.002

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., & Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hillen, J. (2019, 11). Web scraping for food price research. *British Food Journal*, *121*, 3350–3361. doi: 10.1108/BFJ-02-2019-0081

Hochreiter, S., & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735

Ismail, F. N., Woodford, B. J., & Licorish, S. A. (2019). Evaluating the boundaries of big data environments for machine learning [Conference paper]. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11919 LNAI*, 253 – 264. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076535307&doi=10.1007%2f978-3-030-35288-2_21&partnerID=40&md5=699ce2d59b11f040a7386c7ffc674625 (Cited by: 0) doi: 10.1007/978-3-030-35288-2_21

Ivanov, T., & Pergolesi, M. (2020, 3). The impact of columnar file formats on sql-on-hadoop engine performance: A study on orc and parquet. *Concurrency and Computation: Practice and Experience*, *32*. doi: 10.1002/cpe.5523

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. Retrieved from https://arxiv.org/abs/2109.01084 doi: 10.48550/ARXIV.2109.01084

Kanchan, S., Kaur, P., & Apoorva, P. (2021). Empirical evaluation of nosql and relational database systems [Article]. *Recent Advances in Computer Science and Communications*, *14*(8), 2637 – 2650. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123515196&doi=10.2174%2f2666255813999200612113208&partnerID=40&md5=66fa4cb888ba53cca4593043527037cb (Cited by: 1) doi: 10.2174/2666255813999200612113208

Kaur, K., & Sachdeva, M. (2017). Performance evaluation of newsql databases. In *2017 international conference on inventive systems and control (icisc)* (p. 1-5). doi: 10.1109/ICISC.2017.8068585

Kausar, M. A., & Nasar, M. (2021). Sql versus nosql databases to assess their appropriateness for big data application [Article]. *Recent Advances in Computer Science and Communications*, *14*(4), 1098 – 1108. (Cited by: 4) doi: 10.2174/2213275912666191028111632

Kertkeidkachorn, N., & Ichise, R. (2020). Pmap: Ensemble pre-training models for product matching. In *Mwpd@ iswc.*

Khan, R. Z. (2015, 06). Distributed computing: An overview. *Int. J. Advanced Networking and Applications*, *07*, 2630-2635.

Klik, M. (2022). fst: Lightning fast serialization of data frames [Computer software manual]. Retrieved from http://www.fstpackage.org (R package version 0.9.6)

Kornacker, M., Behm, A., Bittorf, V., Bobrovytsky, T., Ching, C., Choi, A., … Yoder, M. (2015). Impala: A modern, open-source sql engine for hadoop [Conference paper]. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084011754&partnerID=40&md5=d63b7b15f0aa3a401fada1a6d67f33fa (Cited by: 220)

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the hadoop ecosystem [Article]. *Journal of Big Data*, *2*(1). Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85013974691&doi=10.1186%2fs40537-015-0032-1&partnerID=40&md5=a257aecdd96c8ea16435354e892711ae (Cited by: 304; All Open Access, Gold Open Access) doi: 10.1186/s40537-015-0032-1

Lee, J., & Dernoncourt, F. (2016, 03). Sequential short-text classification with recurrent and convolutional neural networks. In (p. 515-520). doi: 10.18653/v1/N16-1062

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.* arXiv. Retrieved from https://arxiv.org/abs/1907.11692 doi: 10.48550/ARXIV.1907.11692

Ma, S., Yang, J., Huang, H., Chi, Z., Dong, L., Zhang, D., … Wei, F. (2020). Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *ArXiv*, *abs/2012.15547*.

Macias, P., Stelmasiak, D., & Szafranek, K. (2022). Nowcasting food inflation with a massive amount of online prices. *International Journal of Forecasting*. doi: 10.1016/j.ijforecast.2022.02.007

Mahajan, K., & Tomar, S. (2021). Covid-19 and supply chain disruption: Evidence from food markets in india. *American Journal of Agricultural Economics*, *103*(1), 35-52. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/ajae.12158 doi: https://doi.org/10.1111/ajae.12158

Martindale, H., Rowland, E., Flower, T., & Clews, G. (2020). Semi-supervised machine learning with word embedding for classification in price statistics. *Data & Policy*, *2*, e12. doi: 10.1017/dap.2020.13

National Academies of Sciences, Engineering, and Medicine. (2022). *Modernizing the consumer price index for the 21st century.* doi: https://doi.org/10.17226/26485

Pavlo, A., & Aslett, M. (2016, sep). What's really new with newsql? *SIGMOD Rec.*, *45*(2), 45–55. Retrieved from https://doi.org/10.1145/3003665.3003674 doi: 10.1145/3003665.3003674

Qi, Z. (2020). The text classification of theft crime based on tf-idf and xgboost model. In *2020 ieee international conference on artificial intelligence and computer applications (icaica)* (p. 1241-1246). doi: 10.1109/ICAICA50127.2020.9182555

Raasveldt, M., & Mühleisen, H. (2020). Data management for data science - towards embedded analytics. In *Conference on innovative data systems research.*

Raasveldt, M., & Mühleisen, H. (2019). Duckdb: An embeddable analytical database [Conference paper]. In (p. 1981 – 1984). (Cited by: 30; All Open Access, Green Open Access) doi: 10.1145/3299869.3320212

Santana, I. N., de OLIVEIRA, R. S., & NASCIMENTO, E. G. S. (n.d.). Text classification of news using deep learning and natural language processing models based on.

Seo, S., Kim, C., Kim, H., Mo, K., & Kang, P. (2020). Comparative study of deep learning-based sentiment classification. *IEEE Access*, *8*, 6861-6875. doi: 10.1109/ACCESS.2019.2963426

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The hadoop distributed file system [Conference paper]. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957838299&doi=10.1109%2fMSST.2010.5496972&partnerID=40&md5=964ca2584fceb98153b2ff0a0f3b3e7d (Cited by: 3527) doi: 10.1109/MSST.2010.5496972

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications [Conference paper]. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85072924634&doi=10.1109%2fICCCI.2019.8821809&partnerID=40&md5=53d6555bbff49891567e27b0eadec3cb (Cited by: 22) doi: 10.1109/ICCCI.2019.8821809

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.

Souza, F., & Souza Filho, J. (2022, 12). *Embedding generation for text classification of brazilian portuguese user reviews: from bag-of-words to transformers.*

Tauro, C. J. M., Patil, B. R., & Prashanth, K. R. (2013). A comparative analysis of different nosql databases on data model, query model and replication model..

The Apache Software Foundation. (2019). *Apache arrow project.* Retrieved from https://arrow.apache.org/overview/

Topol, M. (2022). *In-memory analytics with apache arrow.* UK: Packt Publishing.

United Nations Statistics Division, U. (2018). Classification of individual consumption according to purpose (coicop) [Computer software manual]. Retrieved from https://unstats.un.org/unsd/classifications/business-trade/desc/COICOP_english/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf

Ursa Labs. (2019). *Columnar file performance check-in for python and r: Parquet, feather, and fst.* Ursa Labs. Retrieved from https://ursalabs.org/blog/2019-10-columnar-perf/

Uyanga, S., Munkhtsetseg, N., Batbayar, S., & Bat-Ulzii, S. (2021). A comparative study of nosql and relational database [Conference paper]. *Smart Innovation, Systems and Technologies*, *212*, 116 – 122. (Cited by: 1) doi: 10.1007/978-981-33-6757-9_16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., … Baldeschwieler, E. (2013). Apache hadoop yarn: Yet another resource negotiator [Conference paper]. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84893249524&doi=10.1145%2f2523616.2523633&partnerID=40&md5=5794113a8b27dc18de2acdf7dff629f8 (Cited by: 1322) doi: 10.1145/2523616.2523633

Wang, J., Wang, Z., Zhang, D., & Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. In *Proceedings of the 26th international joint conference on artificial intelligence* (p. 2915–2921). AAAI Press.

Wang, R., & Yang, Z.-M. (2017). Sql vs nosql: A performance comparison..

White, T. (2015). *Hadoop: The definitive guide.* e United States of America: O'Reilly Media.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.6 doi: 10.18653/v1/2020.emnlp-demos.6

Xiao, L., Wang, G., & Zuo, Y. (2018). Research on patent text classification based on word2vec and lstm. In *2018 11th international symposium on computational intelligence and design (iscid)* (Vol. 01, p. 71-74). doi: 10.1109/ISCID.2018.00023

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets [Conference paper]. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085251984&partnerID=40&md5=642088c91d2042d23b52ba4f50e0bc20 (Cited by: 3711)

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., … Stoica, I. (2016, oct). Apache spark: A unified engine for big data processing. *Commun. ACM*, *59*(11), 56–65. Retrieved from https://doi.org/10.1145/2934664 doi: 10.1145/2934664

Zahera, H. M., & Sherif, M. (2020). Probert: Product data classification with fine-tuning bert model. In *Mwpd@ iswc.*

Zhang, Z., Bizer, C., Peeters, R., & Primpeli, A. (2020). Mwpd2020: Semantic web challenge on mining the web of html-embedded product data. In *Mwpd@iswc.*

Appendix A

ECOICOP Definition

Code	PT Description	EN Description	Includes	Excludes
01.1.1.1	Arroz	Rice	- rice in all forms, including rice prepared with meat, fish, seafood or vegetables (if rice represents the most important part)	- rice flour (01.1.1.2)
01.1.1.2	Farinhas e outros cereais	Flours and other cereals	- maize, wheat, barley, oats, rye and other cereals in the form of grain, flour or meal - rice flour	
01.1.1.3	Pão	Bread	- bread and bread rolls	
01.1.1.4	Outros produtos de padaria	Other bakery products	- crisp bread, rusks, toasted bread, biscuits, gingerbread, wafers, waffles, crumpets, muffins, croissants, cakes, tarts, sweet pies	- meat pies (01.1.2.8) - fish pies (01.1.3.6)
01.1.1.5	Pizzas e quiches	Pizza and quiche	- farinaceous-based (flour based) products prepared with meat, fish, seafood, cheese, vegetables or fruit	- meat pies (01.1.2.8) - fish pies (01.1.3.6)
01.1.1.6	Massas alimentícias e cuscuz	Pasta products and couscous	- ravioli, cannelloni, lasagne etc. - pasta products in all forms - couscous, including prepared dishes made basically of pasta or couscous	
01.1.1.7	Cereais para pequeno-almoço	Breakfast cereals	- cornflakes, oatflakes, muesli etc.	
01.1.1.8	Outros produtos à base de cereais	Other cereal products	- mixes and doughs for the preparation of bakery products - other cereal products (malt, malt flour, malt extract, potato starch, tapioca, sago and other starches)	

Code	PT Description	EN Description	Includes	Excludes
01.1.2.1	Carne de bovino	Beef and veal	- fresh, chilled or frozen meat of bovine animals - cow or veal purchased live for consumption as food - minced meat made of beef or veal	
01.1.2.2	Carne de suíno	Pork	- fresh, chilled or frozen meat of swine - pork purchased live for consumption as food - minced meat made of pork	
01.1.2.3	Carne de ovino e caprino	Lamb and goat	- fresh, chilled or frozen meat of sheep and goat - lamb and goat purchased live for consumption as food Also - minced meat made of lamb and goat	
01.1.2.4	Aves de capoeira	Poultry	- fresh, chilled or frozen meat of poultry (chicken, duck, goose, turkey, guinea fowl etc.) - poultry purchased live for consumption as food Also - minced meat made of poultry	

Code	PT Description	EN Description	Includes	Excludes
01.1.2.5	Outras carnes	Other meat	Includes fresh, chilled or frozen meat of: - horse, mule, donkey, camel and the like - hare, rabbit and game (antelope, deer, boar, pheasant, grouse, pigeon, quail, etc.) - meat of marine mammals (seals, walruses, whales, etc.) and exotic animals (kangaroo, ostrich, alligator, etc.) - animals purchased live for consumption as food	
01.1.2.6	Miudezas comestíveis	Edible offal	- fresh, chilled, smoked or frozen edible offal	
01.1.2.7	Carne seca, salgada ou fumada	Dried, salted or smoked meat	- dried, salted or smoked meat (sausages, salami, bacon, ham, etc.)	
01.1.2.8	Outras preparações à base de carne	Other meat preparations	- other preserved or processed meat and meat-based preparations (canned meat, meat extracts, meat juices, meat pies, etc.) - dumplings, filled pancakes with meat - minced meat, raw or prepared if mixed meat from more than one kind of minced meat - all kind of pâté, including liver pâté	- pizza and quiche (01.1.1.5)
01.1.3.1	Peixe fresco ou refrigerado	Fresh or chilled fish	- fresh or chilled fish - fish purchased live for consumption as food	
01.1.3.2	Peixe congelado	Frozen fish	- frozen fish	

Code	PT Description	EN Description	Includes	Excludes
01.1.3.3	Marisco fresco ou refrigerado	Fresh or chilled seafood	- fresh or chilled seafood (crustaceans, molluscs and other shellfish, sea snails) - seafood purchased live for consumption as food Also - land crabs, land snails and frogs	
01.1.3.4	Marisco congelado	Frozen seafood	- frozen seafood	
01.1.3.5	Peixe e marisco seco, fumado ou salgado	Dried, smoked or salted fish and seafood	- dried, smoked or salted fish and seafood	
01.1.3.6	Outras preparações à base de peixe e marisco transformado ou conservado	Other preserved or processed fish and seafood-based preparations	- other preserved or processed fish and seafood and fish and seafood-based preparations (canned fish and seafood, caviar and other hard roes, fish pies, battered fish, etc.) - dumplings, filled pancakes with fish	- pizza and quiche containing fish and seafood (01.1.1.5) - soups, broths and stocks containing fish and seafood (01.1.9.9)
01.1.4.1	Leite gordo fresco	Fresh whole milk	- fresh, whole milk - pasteurized or sterilized milk Also - ”ultra-pasteurized” or ”UHT” milk	
01.1.4.2	Leite magro fresco	Fresh low fat milk	- fresh, low fat milk - pasteurized or sterilized milk Also - ”ultra-pasteurized” or ”UHT” milk, semi-skimmed milk and skimmed milk	
01.1.4.3	Leite conservado	Preserved milk	- condensed, evaporated or powdered milk	
01.1.4.4	Iogurte	Yoghurt	- yoghurt containing or not sugar, cocoa, fruit or flavourings	

Code	PT Description	EN Description	Includes	Excludes
01.1.4.5	Queijos e requeijão	Cheese and curd	- hard, semi-hard, blue cheese, cottage cheese, mozzarella, ””fromage blanc””	
01.1.4.6	Outros produtos lácteos	Other milk products	- cream, milk-based desserts, milk-based beverages and other similar milk-based products - dairy products not based on milk such as soya milk	- butter and butter products (01.1.5.1) - soya based desserts (01.1.9.4)
01.1.4.7	Ovos	Eggs	- eggs and egg products made wholly from eggs	
01.1.5.1	Manteiga	Butter	- butter and butter products (butter oil, ghee, etc.)	
01.1.5.2	Margarina e outras gorduras vegetais	Margarine and other vegetable fats	- margarine (including diet margarine) and other vegetable fats (including peanut butter)	
01.1.5.3	Azeite	Olive oil	- olive oil	
01.1.5.4	Outros óleos alimentares	Other edible oils	- other edible oils (corn oil, sunflower-seed oil, cottonseed oil, soybean oil, groundnut oil, walnut oil, etc.)	- cod or halibut liver oil (06.1.1.0)
01.1.5.5	Outras gorduras animais comestíveis	Other edible animal fats	- edible animal fats (lard, etc.)	
01.1.6.1	Fruta fresca ou refrigerada	Fresh or chilled fruit	- melons and water melons, berries	
01.1.6.2	Fruta congelada	Frozen fruit	- frozen fruit - frozen berries	
01.1.6.3	Frutos secos e frutos de casca rija	Dried fruit and nuts	- dried fruit, fruit peel, fruit kernels - nuts and edible seeds - dry berries	

Code	PT Description	EN Description	Includes	Excludes
01.1.6.4	Frutas em conserva e produtos à base de frutas em conserva	Preserved fruit and fruit-based products	- preserved fruit and fruit-based products - dietary preparations and culinary ingredients based exclusively on fruit - canned or tinned fruit	
01.1.7.1	Produtos hortícolas frescos ou refrigerados, exceto batatas e outros tubérculos	Fresh or chilled vegetables other than potatoes and other tubers	- fresh or chilled vegetables cultivated for their leaves or stalks (asparagus, broccoli, cauliflower, endives, fennel, spinach, etc.), for their fruit (aubergines, cucumbers, courgettes, green peppers, pumpkins, tomatoes, etc.) and for their roots (beetroots, carrots, onions, parsnips, radishes, turnips, etc.)	
01.1.7.2	Produtos hortícolas congelados, exceto batatas e outros tubérculos	Frozen vegetables other than potatoes and other tubers	- frozen vegetables cultivated for their leaves or stalks (asparagus, broccoli, cauliflower, endives, fennel, spinach, etc.), for their fruit (aubergines, cucumbers, courgettes, green peppers, pumpkins, tomatoes, etc.) and for their roots (beetroots, carrots, onions, parsnips, radishes, turnips, etc.)	

Code	PT Description	EN Description	Includes	Excludes
01.1.7.3	Produtos hortícolas secos, outros produtos hortícolas conservados ou transformados	Dried vegetables, other preserved or processed vegetables	- vegetable-based products, dietary preparations and culinary ingredients based exclusively on vegetables - mixtures of vegetables - canned or tinned vegetables - pulses	
01.1.7.4	Batatas	Potatoes	- fresh, chilled and preserved potatoes Also - frozen preparations such as chipped potatoes	- potato starch (01.1.1.8) - sweet potatoes (01.1.7.6)
01.1.7.5	Batatas fritas	Crisps	- potato crisps (simple or using various flavourings and ingredients including seasonings, herbs, spices, cheeses, and artificial additives) - crisps made from potato, but may also be made from corn, maize, tapioca and other tuber	
01.1.7.6	Outros tubérculos e produtos de tubérculos	Other tubers and products of tuber vegetables	- manioc, arrowroot, cassava, sweet potatoes, etc. - products of tuber vegetables (flours, meals, flakes, purées)	
01.1.8.1	Açúcar	Sugar	- cane or beet sugar, unrefined or refined, powdered, crystallized or in lumps	
01.1.8.2	Doces de fruta, doces de citrinos e mel	Jams, marmalades and honey	- jams, marmalades, compotes, jellies, fruit purées and pastes, natural and artificial honey, maple syrup, molasses and parts of plants preserved in sugar	

Code	PT Description	EN Description	Includes	Excludes
01.1.8.3	Chocolate	Chocolate	- chocolate and cocoa-based foods and cocoa-based dessert preparations	- cocoa and chocolate-based powder (01.2.1.3)
01.1.8.4	Produtos de confeitaria	Confectionery products	- chewing gum, sweets, toffees, pastilles and other confectionery products	
01.1.8.5	Gelo comestível e gelados	Edible ices and ice cream	- sorbet	
01.1.8.6	Sucedâneos artificiais do açúcar	Artificial sugar substitutes	- artificial sugar substitutes	
01.1.9.1	Molhos, condimentos	Sauces, condiments	- sauces, condiments, seasonings (mustard, mayonnaise, ketchup, soy sauce, etc.), vinegar	
01.1.9.2	Sal, especiarias e ervas aromáticas	Salt, spices and culinary herbs	- salt, spices (pepper, pimento, ginger, etc.), culinary herbs (parsley, rosemary, thyme, etc.)	
01.1.9.3	Alimentos para bebés	Baby food	- homogenized baby food irrespective of the composition	
01.1.9.4	Pratos preparados	Ready-made meals	- ready-to-eat dishes (tinned food, frozen food or meals prepared in the day), irrespective of the composition, are classified in this category when the price only covers the cost of the product - sandwiches	- pizza and quiche (01.1.1.5) - pasta and couscous prepared in all forms (01.1.1.6) - meat pies (01.1.2.8) - fish pies (01.1.3.6)

Code	PT Description	EN Description	Includes	Excludes
01.1.9.9	Outros produtos alimentares, n.e.	Other food products n.e.c.	- prepared baking powders, baker's yeast, dessert preparations, soups, broths, stocks, culinary ingredients, etc. - dietary preparations irrespective of the composition	- diet margarine (01.1.5.2) - dietary preparations based exclusively on fruit (01.1.6.4)
01.2.1.1	Café	Coffee	- coffee, whether or not decaffeinated, roasted or ground, including instant coffee Also - coffee substitutes - extracts and essences of coffee	
01.2.1.2	Chá	Tea	- tea, maté and other plant products for infusions Also - tea substitutes - extracts and essences of tea	
01.2.1.3	Cacau e chocolate em pó	Cocoa and powdered chocolate	- cocoa, whether or not sweetened, and chocolate-based powder Also - cocoa-based beverage preparations	- chocolate in bars or slabs (01.1.8.3) - cocoa-based food and cocoa-based dessert preparations (01.1.8.3)
01.2.2.1	Água mineral ou água de nascente	Mineral or spring waters	- mineral or spring waters - all drinking water sold in containers	
01.2.2.2	Refrigerantes	Soft drinks	- soft drinks such as sodas, lemonades and colas	- non-alcoholic beverages which are generally alcoholic such as non-alcoholic beer (02.1)

Code	PT Description	EN Description	Includes	Excludes
01.2.2.3	Sumos de fruta e de produtos hortícolas	Fruit and vegetable juices	- fruit and vegetable juices - syrups and concentrates for the preparation of beverages	
02.1.1.1	Bebidas espirituosas e licores	Spirits and liqueurs	- eaux-de-vie, liqueurs and other spirits with high alcohol content Also - mead - aperitifs other than wine-based aperitifs	Excludes: - wine-based aperitifs (02.1.2.4)
02.1.1.2	Refrigerantes com álcool (alcopops)	Alcoholic soft drinks	- soda-water types with a low alcohol content	
02.1.2.1	Vinhos de uva	Wine from grapes	- champagne and other sparkling wines	
02.1.2.2	Vinhos de outros frutos	Wine from other fruits	- cider and perry, including sake	
02.1.2.3	Vinhos enriquecidos com álcool	Fortified wines	- vermouth, sherry, port wine	
02.1.2.4	Bebidas à base de vinho	Wine-based drinks	- wine-based aperitifs, non-alcoholic wine	
02.1.3.1	Cerveja tipo lager	Lager beer	- Pilsner, Bock, Dortmunder Export and Märzen lager beers Also - pale lager and dark lagers, such as Dunkel and Schwarzbier	
02.1.3.2	Outro tipo de cerveja com álcool	Other alcoholic beer	- Ale beers Also - hybrid or mixed style beers, such as Altbier and Kölsch, steam beers, fruit and vegetable beers, herb and spiced beers, wood-aged beers, smoked beers or champagne style beers	

Code	PT Description	EN Description	Includes	Excludes
02.1.3.3	Cerveja de baixo teor alcoólico ou não alcoólica	Low and non-alcoholic beer	- low alcoholic beer do not have a harmonized % ABV in all EU Member States but it should be around less than 1% ABV	
02.1.3.4	Bebidas à base de cerveja	Beer-based drinks	- beer with soda ("panache"), beer with Coca-Cola Also - "shandy" (mix of beer and soda-water with ginger taste)	

Table A.1: Description of ECOICOP Categories: Food and Beverage

Appendix B

Distribution of products per category

Category	**Supermarket A**	**Supermarket A + B**	**All Supermarkets**
Non-Food and Non-Beverage	60.058%	53.887%	47.887%
Wine from grapes	3.801%	4.716%	4.552%
Other bakery products	3.124%	3.700%	4.308%
Chocolate	1.835%	2.500%	2.462%
Yoghurt	1.562%	1.884%	2.023%
Cheese and curd	1.325%	1.432%	1.839%
Sauces, condiments	1.267%	1.569%	1.794%
Dried vegetables, other preserved or processed vegetables	1.296%	1.376%	1.666%
Dried, salted or smoked meat	1.111%	1.360%	1.661%
Coffee	1.242%	1.334%	1.573%
Pasta products and couscous	1.398%	1.357%	1.426%
Outros produtos lácteos	1.005%	1.113%	1.393%
Confectionery products	1.121%	1.442%	1.324%
Soft drinks	0.936%	1.115%	1.314%
Jams, marmalades and honey	0.932%	1.155%	1.213%
Fruit and vegetable juices	0.892%	1.115%	1.195%
Fresh or chilled vegetables other than potatoes and other tubers	0.812%	0.887%	1.149%
Tea	0.725%	0.875%	1.110%
Salt, spices and culinary herbs	0.757%	0.910%	1.078%
Other preserved or processed fish and seafood-based preparations	0.819%	0.953%	1.076%
Baby food	0.987%	1.073%	1.040%
Spirits and liqueurs	0.717%	0.929%	1.038%
Ready-made meals	0.816%	0.684%	1.005%
Crisps	0.746%	0.760%	0.948%
Breakfast cereals	0.637%	0.786%	0.909%
Dried fruit and nuts	0.659%	0.847%	0.886%
Edible ices and ice cream	0.597%	0.732%	0.815%
Flours and other cereals	0.561%	0.623%	0.782%
Other cereal products	0.612%	0.692%	0.717%
Bread	0.473%	0.525%	0.701%
Frozen fish	0.499%	0.529%	0.609%

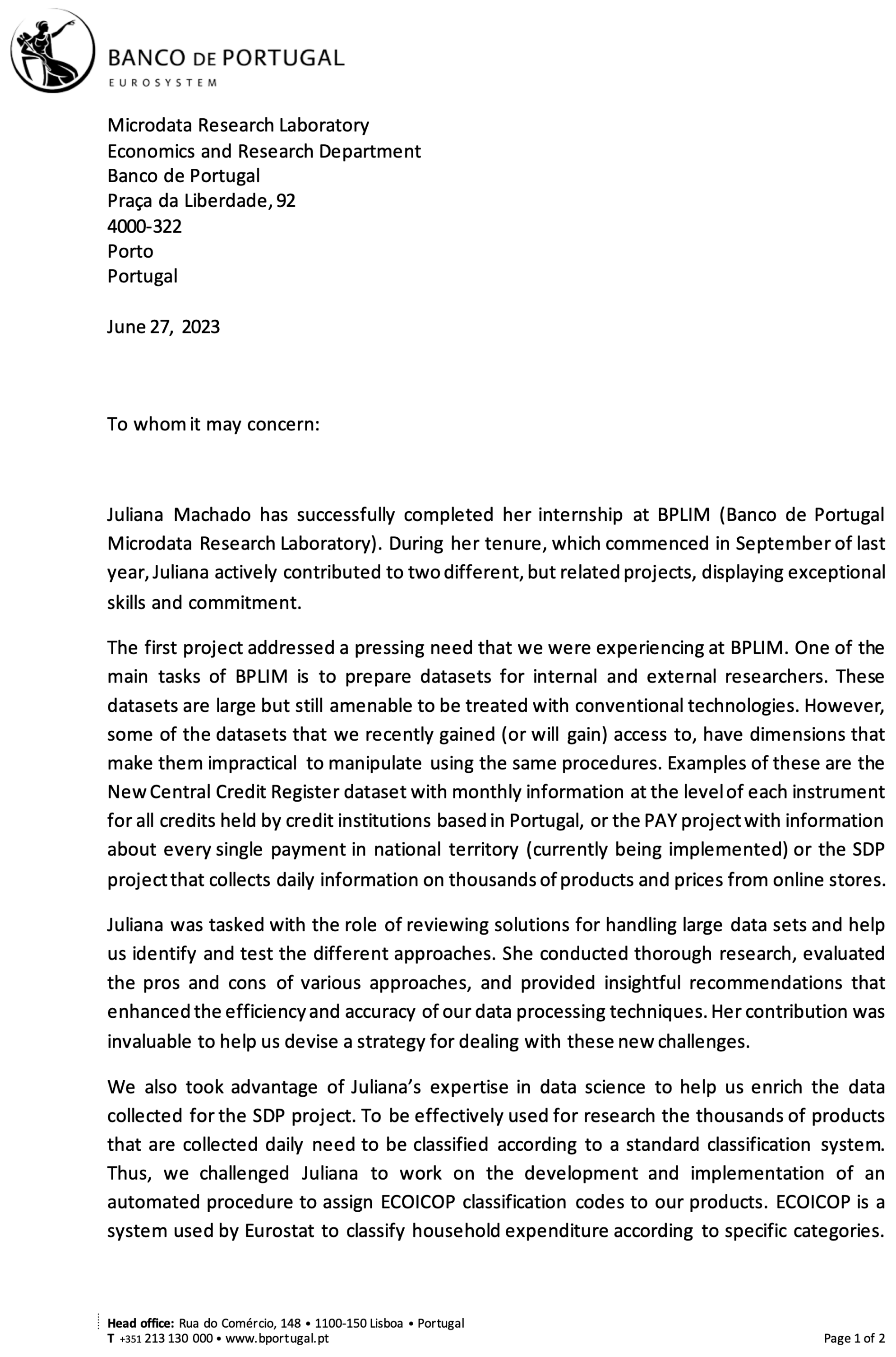
Category	Supermarket A	Supermarket A + B	All Supermarkets
Other food products n.e.c.	0.411%	0.517%	0.584%
Other meat preparations	0.444%	0.447%	0.536%
Fresh or chilled fruit	0.335%	0.369%	0.494%
Mineral or spring waters	0.328%	0.388%	0.492%
Lager beer	0.390%	0.423%	0.487%
Rice	0.339%	0.329%	0.410%
Fortified wines	0.404%	0.350%	0.390%
Pizza and quiche	0.237%	0.289%	0.366%
Frozen seafood	0.375%	0.395%	0.364%
Margarine and other vegetable fats	0.288%	0.320%	0.350%
Olive oil	0.346%	0.329%	0.349%
Fresh low fat milk	0.197%	0.263%	0.303%
Frozen vegetables other than potatoes and other tubers	0.131%	0.193%	0.265%
Preserved fruit and fruit-based products	0.182%	0.214%	0.264%
Poultry	0.200%	0.200%	0.232%
Fresh or chilled fish	0.175%	0.148%	0.227%
Potatoes	0.131%	0.153%	0.209%
Beef and veal	0.222%	0.188%	0.203%
Other alcoholic beer	0.127%	0.151%	0.189%
Pork	0.193%	0.169%	0.188%
Butter	0.135%	0.146%	0.180%
Sugar	0.084%	0.115%	0.171%
Cocoa and powdered chocolate	0.124%	0.125%	0.154%
Fresh or chilled seafood	0.102%	0.103%	0.143%
Other edible oils	0.120%	0.103%	0.125%
Wine from other fruits	0.069%	0.075%	0.100%
Eggs	0.091%	0.080%	0.094%
Dried, smoked or salted fish and seafood	0.127%	0.094%	0.090%
Artificial sugar substitutes	0.062%	0.071%	0.085%
Wine-based drinks	0.047%	0.071%	0.080%

Category	**Supermarket A**	**Supermarket A + B**	**All Supermarkets**
Low and non-alcoholic beer	0.044%	0.059%	0.068%
Other milk products	0.051%	0.059%	0.068%
Other tubers and products of tuber vegetables	0.044%	0.056%	0.055%
Frozen fruit	0.036%	0.042%	0.049%
Alcoholic soft drinks	0.029%	0.035%	0.031%
Other meat	0.018%	0.016%	0.029%
Lamb and goat	0.000%	0.009%	0.027%
Leite gordo fresco	0.011%	0.016%	0.020%
Edible offal	0.018%	0.016%	0.020%
Other edible animal fats	0.007%	0.007%	0.010%
Beer-based drinks	0.004%	0.005%	0.006%

Table B.1: Distribution of products per category

Appendix C

Banco de Portugal Letter



BANCO DE PORTUGAL
EUROSYSTEM

Microdata Research Laboratory
Economics and Research Department
Banco de Portugal
Praça da Liberdade, 92
4000-322
Porto
Portugal

June 27, 2023

To whom it may concern:

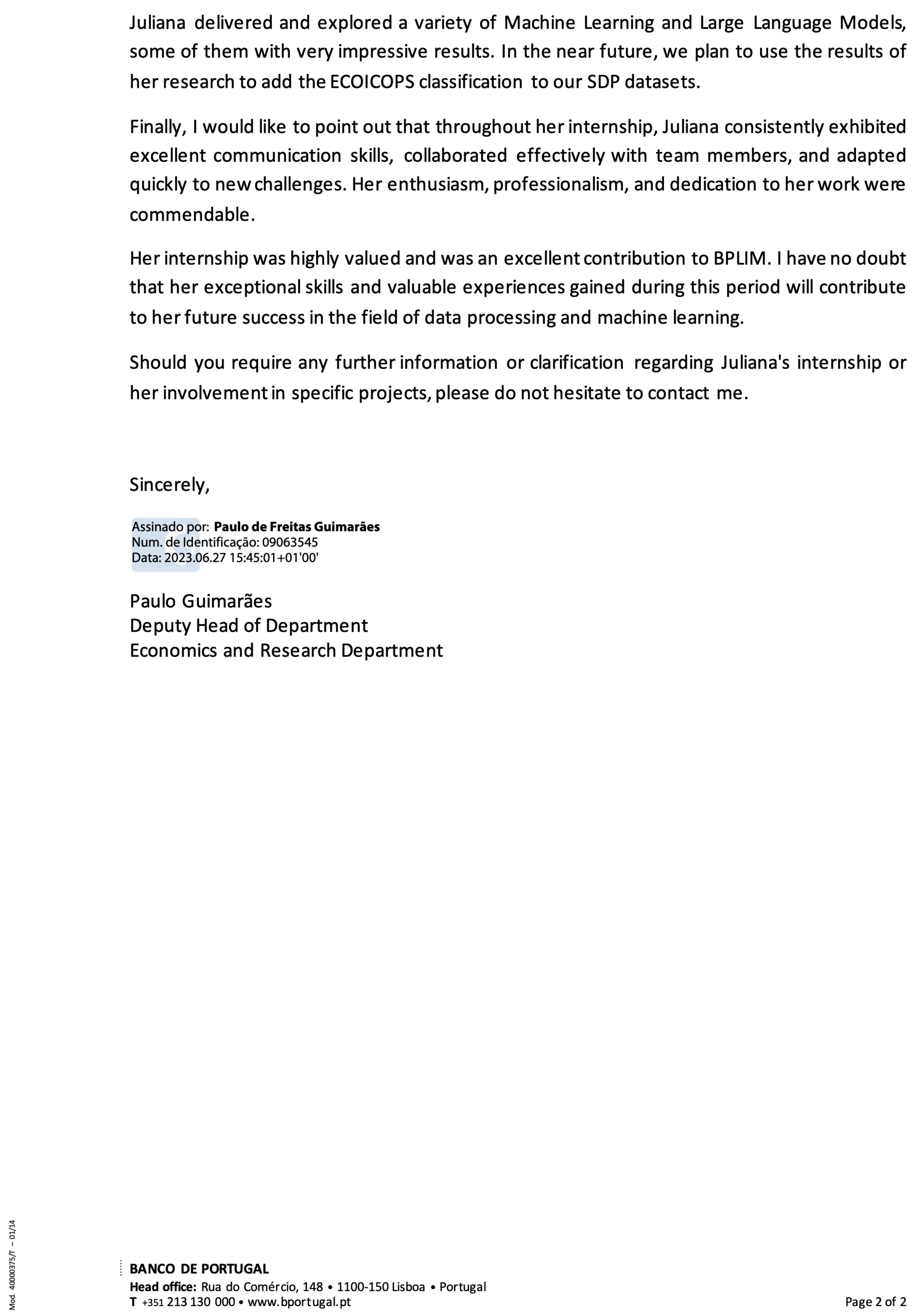
Juliana Machado has successfully completed her internship at BPLIM (Banco de Portugal Microdata Research Laboratory). During her tenure, which commenced in September of last year, Juliana actively contributed to two different, but related projects, displaying exceptional skills and commitment.

The first project addressed a pressing need that we were experiencing at BPLIM. One of the main tasks of BPLIM is to prepare datasets for internal and external researchers. These datasets are large but still amenable to be treated with conventional technologies. However, some of the datasets that we recently gained (or will gain) access to, have dimensions that make them impractical to manipulate using the same procedures. Examples of these are the New Central Credit Register dataset with monthly information at the level of each instrument for all credits held by credit institutions based in Portugal, or the PAY project with information about every single payment in national territory (currently being implemented) or the SDP project that collects daily information on thousands of products and prices from online stores.

Juliana was tasked with the role of reviewing solutions for handling large data sets and help us identify and test the different approaches. She conducted thorough research, evaluated the pros and cons of various approaches, and provided insightful recommendations that enhanced the efficiency and accuracy of our data processing techniques. Her contribution was invaluable to help us devise a strategy for dealing with these new challenges.

We also took advantage of Juliana's expertise in data science to help us enrich the data collected for the SDP project. To be effectively used for research the thousands of products that are collected daily need to be classified according to a standard classification system. Thus, we challenged Juliana to work on the development and implementation of an automated procedure to assign ECOICOP classification codes to our products. ECOICOP is a system used by Eurostat to classify household expenditure according to specific categories.

Mod. 40000375/T – 01/14



Juliana delivered and explored a variety of Machine Learning and Large Language Models, some of them with very impressive results. In the near future, we plan to use the results of her research to add the ECOICOPS classification to our SDP datasets.

Finally, I would like to point out that throughout her internship, Juliana consistently exhibited excellent communication skills, collaborated effectively with team members, and adapted quickly to new challenges. Her enthusiasm, professionalism, and dedication to her work were commendable.

Her internship was highly valued and was an excellent contribution to BPLIM. I have no doubt that her exceptional skills and valuable experiences gained during this period will contribute to her future success in the field of data processing and machine learning.

Should you require any further information or clarification regarding Juliana's internship or her involvement in specific projects, please do not hesitate to contact me.

Sincerely,

Assinado por: **Paulo de Freitas Guimarães**
Num. de Identificação: 09063545
Data: 2023.06.27 15:45:01+01'00'

Paulo Guimarães
Deputy Head of Department
Economics and Research Department