Recommender sytem based on NLP: a support tool for the publishing sector

Jessie Caridad Martín Sujo

http://hdl.handle.net/10803/688979

Data de defensa: 26-07-2023

DOCTORAL THESIS

Title	Recommender system based on NLP: a support tool for the publishing sector
Presented by	Jessie Caridad Martín Sujo
Centre	La Salle Digital Engineering School
Department	Engineering
Directed by	Dra. Elisabet Golobardes Ribé

Dedication

To my four driving forces: my mother, my grandmother, my boyfriend and my grandfather (who watches over me from heaven).

Declaration

I **Jessie Caridad Martín Sujo**, declare that the ideas, judgments, evaluations, interpretations, bibliographical consultations, definitions and conceptualizations exposed in this work; as well as, the procedures and tools used in the investigation, are the absolute responsibility of the author of the curricular integration work. Likewise, I accept the internal regulations of the Ramon Llull University.

Barcelona, May 2023.

Jessie Caridad Martín Sujo

Acknowledgment

This doctoral thesis research has been the result of numerous experiences, consultations and sleeplessness. There are many people (an endless list) to whom I owe a great debt of gratitude for their invaluable support, comments, critiques, corrections, and references. All of them undoubtedly furthered my experiences during these three years of arduous, challenging, and sometimes lonely research.

Although it is true that happiness can be found even in the darkest moments, if one only remembers to turn on the light... I want to make a special dedication to my boyfriend, who has found a way to turn on that little light in the middle of my darkness. Thank you my love for your unconditional support during these years. And recognize that this achievement would not have been possible without the strength and power that my mother and grandmother transmit to me from 7152 km.

Also, I would like to extend my heartfelt thanks to "Ajeev", whose music inspires us to rise up and expand beyond our imaginary limits.

Finally, and no less important (quite the opposite), I would like to thank all my friends from Seville, Barcelona and Burgos for listening to me in my moments of anguish, for being my confidants and for these years of true friendship.

Abstract

In recent years, the Spanish publishing industry has been getting closer and closer to digital transformation, however there are still many challenges to overcome, especially in recommendations to the end customer (readers). With the advancement of Machine learning, a branch of Artificial Intelligence, it is possible that many of these challenges can be met.

However, it is important to note that the books will have to be recommended by a literary genre individually, since they present completely different characteristics from each other. For the analysis of Non-fiction books, this work will be based on historical sales and the author's influence on social networks. That is why, at the architecture level, it will be defined as a boosting model, which will reduce errors in predictive data analysis. And for the analysis of Fiction books, it will be based on the emotions that the text transmits. That is why, at the architecture level, it will be determined by Transformers networks, whose responsibility will be to manipulate sequential data, mainly in the field of NLP. Finally, and at the design level, the outputs of these subsystems will serve as input data for the recommendation engine.

In this thesis, a final recommendation system based on Natural Language Processing (NLP) is proposed, whose main contributions are: (1) the integration of multidisciplinary professions such as psychology, literature and artificial intelligence; (2) a design of an intelligent system that recommends based on the behavior of the reader and not only on their purchases; and (3) boost literary quality while still attracting people who don't usually read.

Keywords:
Artificial Intelligence, Machine learning, Natural Language Processing, Transformers, Recommendation Engine, Intelligent System, Publishing Industry.

Resumen

En los últimos años, la industria editorial española se ha ido acercando cada vez más a la transformación digital, sin embargo aún quedan muchos retos por superar, especialmente en las recomendaciones al cliente final (lectores). Con el avance del aprendizaje automático, una rama de la inteligencia artificial, es posible que se puedan cumplir muchos de estos desafíos.

Sin embargo, es importante señalar que los libros tendrán que ser recomendados por un género literario de forma individual, ya que presentan características completamente diferentes entre sí. Para el análisis de los libros de No ficción, este trabajo se basará en las ventas históricas y la influencia del autor en las redes sociales. Por eso, a nivel de arquitectura, se definirá como un modelo potenciador, que reducirá los errores en el análisis predictivo de datos. Y para el análisis de los libros de Ficción, se partirá de las emociones que transmite el texto. Es por ello que, a nivel de arquitectura, estará determinada por las redes de Transformadores, cuya responsabilidad será manipular datos secuenciales, principalmente en el campo de la PNL. Finalmente, y a nivel de diseño, las salidas de estos subsistemas servirán como datos de entrada para el motor de recomendaciones.

En esta tesis se propone un sistema de recomendación final basado en el Procesamiento del Lenguaje Natural (PNL), cuyas principales contribuciones son: (1) la integración de profesiones multidisciplinares como la psicología, la literatura y la inteligencia artificial; (2) un diseño de un sistema inteligente que recomiende con base en el comportamiento del lector y no solo en sus compras; y (3) impulsar la calidad literaria sin dejar de atraer a personas que no suelen leer.

Palabras Clave:
Inteligencia artificial, Machine learning, Procesamiento de lenguaje natural, Transformers, Motor de recomendación,Sistema Inteligente, Industria Editorial.

Contents

List of publications

The results of this thesis are the product of these articles.

- Martín Sujo, J. C., Golobardes i Ribé, E., Vilasís Cardona, X., Jiménez Ruano, V., and Villasmil López, J. (2021, August). Correction to: SmartData: An Intelligent Decision Support System to Predict the Readers Permanence in News. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 2 (pp. C1-C1). Cham: Springer International Publishing. [1]

- Martín Sujo, J. C., Golobardes i Ribé, E., and Vilasís Cardona, X. (2021). CAIT: A Predictive Tool for Supporting the Book Market Operation Using Social Networks. Applied Sciences, 12(1), 366.

- Martín Sujo, J. C., and Golobardes i Ribé, E. Personality Profile of Fictional Characters in Books Using Natural Language Processing. Personality Profile of Fictional Characters in Books Using Natural Language Processing. Submitted to Computer Speech & Language - Elsevier.

- Martín Sujo, J. C., and Golobardes i Ribé, E. Brain l: A book recommender system, 2023. arXiv preprint arXiv:2302.00653. 2023.

[1] Original (with last name error): Sujo, J. C. M., Cardona, X. V., Ruano, V. J., and López, J. V. (2021, September). SmartData: An Intelligent Decision Support System to Predict the Readers Permanence in News. In Proceedings of SAI Intelligent Systems Conference (pp. 326-339). Springer, Cham.

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence.

ANN Artificial Neural Networks.

B5 Big five.

BERT Bidirectional Encoder Representations from Transformers.

BoW Bag of Words.

CBR Case-Based Reasoning.

CF Collaborative Filtering.

ELMo Embeddings from Language Models.

KNN K-Nearest Neighbors.

LSTM Long Short-Term Memory.

MBTI Myers-Briggs Type Indicator.

ML Machine learning.

MLP Multilayer Perceptron.

NER Named Entity Recognition.

NLP Natural Language Processing.

POS Part of speech.

TF-IDF erm Frequency Inverse Document Frequency.

Part I

Initial state of knowledge

Chapter 1

Introduction

"The data is boring, it is the emotion that transcends."
-Anonymous-

Throughout this chapter, the justification and approach to the problem, the starting hypothesis, the main objective, and the specific objectives to be achieved in the development of this doctoral thesis are presented, subsequently demonstrated through development and evaluation of various prototypes. Also the structure of the document is presented.

1.1 Personal background

The doctoral program where the research is framed is the Doctorate in Information Technology and its application in management, architecture and geophysics of La Salle from the Ramón Llull University (URL). This has been carried out in the Research group on Data Science for the Digital Society (DS4DS). The DS4DS is a research group created in 2017 by the Generalitat de Catalunya within the framework of the SGR 2017 Call (2017 SGR 0920) within La Salle. Currently, this group has expanded and is now part of the Smart Society research group created in 2022 by the Generalitat de Catalunya within the framework of the SGR 2022 Call (2022 SGR 01398). Centering its activity around Artificial Intelligence, it stands out in three areas: Large Hadron Collider Beauty Experiment (LHCb), Logistics and Operations, and Natural Language Processing (NLP).

This group is the main coordinator of the Master in Data Science, from which I graduated in 2019. Thanks to the contribution I had within the SmartData project, financed by the Ministry of Economy, Industry and Competitiveness of the Government of Spain and the European Union Fund for Regional Development with the help n^o RTC-2016-5503-7 (MINECO/FEDER, EU), I began to be part of the research team. The main project in which I have worked has been financed by a private company, which we will call "The Editorial" (respecting the anonymity of the medium), maintaining as main technologies: natural language processing and intelligent systems.

Definitely, this research project is a product, in addition to public and private projects, of the learning acquired during the courses offered in the Doctoral program. The thesis has been supervised by Prof. Dra. Elisabet Golobardes i Ribé; and possible thanks to the support and help received from the Smart Society group and La Salle - Ramon Llull University.

To show the evolution during these three years of research for the Doctoral Thesis, below is an infographic of the timeline with all the milestones that will be described during this document.

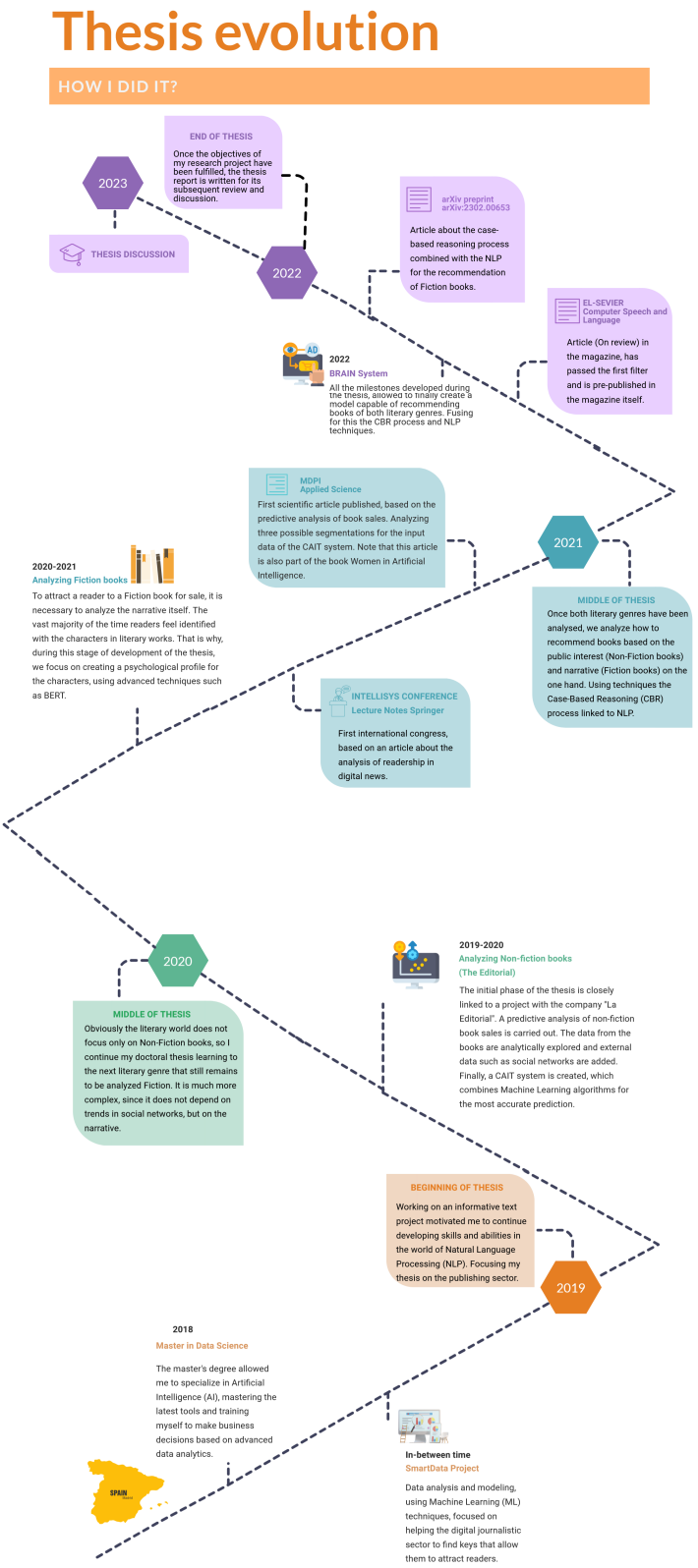

FIGURE 1.1: Timeline of the development process of this doctoral thesis. Source: Own elaboration.

1.2 Research framework

In recent years, the Spanish publishing industry has been seriously affected in the supply of publications that attract readers. Since 2000 to the present, the Federation of Publishers Guilds of Spain (FGEE acronym in Spanish) [1] carries out annual studies in this regard. These have made it possible to monitor the state of health of reading and buying books in the country. It is possible through continuous monitoring of various indicators such as: the proportion of frequent and occasional readers, the number of books purchased and read, the hours dedicated to reading, the adoption of new digital reading habits or the measurement of reading between the children population. It can be seen in Figure 1.2, the evidence of this fact. Until 2019, it is known that 40.3% of Spaniards were immune to the charms of a book [2]. Likewise, there is a relatively significant increase of 2% in sales for 2020 (if the compared to the previous year), due to the fact that reading has been a means of escaping boredom during the COVID19 pandemic.

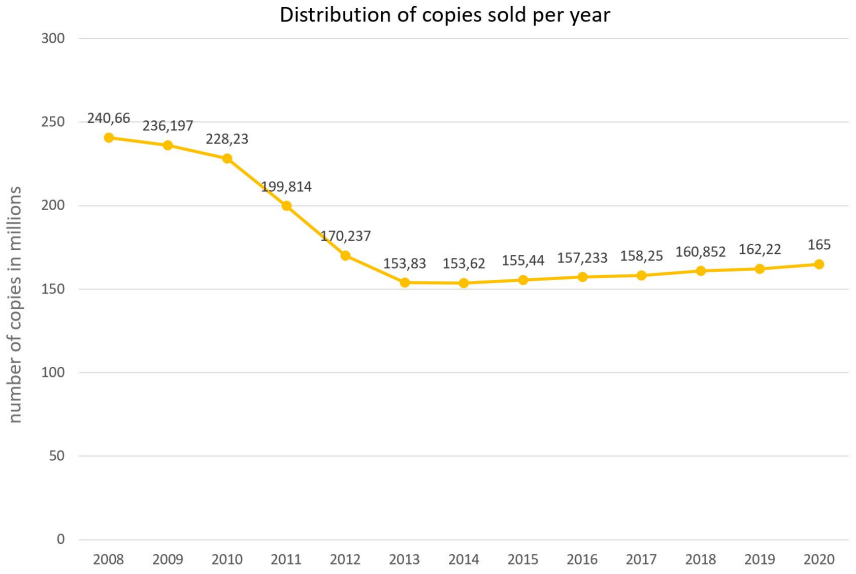

FIGURE 1.2: Sales decline over the years. Source: Own elaboration with data extracted from the federation of publishers.

[1] For further details of this study, consult the statistical data on the official website: https://www.federacioneditores.org/datos-estadisticos .php

[2] For more details on this study, consult the news on the official website: https://www.abc.es/cultura/abci-mas-40-por-ciento-poblacion-solo-papel-201801181107_noticia.html

But why are sales still not reaching the point of splendor they had before 2012?

The consumption of books in Spain is in decline due to multiple factors, even unrelated to those known as the financial crisis, prices, piracy, etc. Possible reasons why fewer books are being sold include:

- Literary critics do not recommend books that present a real affinity with readers (even that may surprise them), but rather focus on the needs of publishers. [1]
- Commercial campaigns are aimed more at selling book volume than at writer-reader interaction. [2]
- Quantity is continuously strengthened more than quality. [3]
- The vertiginous development of the Internet and social networks in the last decade.[4]

A study by Queirós et al. [5], carried out jointly with the Reina Sofía Center for Adolescence and Youth and Fundación MAPFRE, gathers the opinion of 1,401 young people between 14 and 24 years of age and different work groups; that practically half of the young people affirm that, due to the use of social networks, study and reading time decreases. Even recently, systems have been created to analyze the time spent by readers, such as the one implemented in the study by Martín et al. [6], where the authors show an 88% effectiveness in predicting the permanence of readers in digital newspapers. And where the data indicates that the greater the content (length of the text) of the news along with an intermediate number of audios inserted in it, the greater the probability that readers will be attracted to it. Obviously, a book does not have the same characteristics as digital news and audio cannot be inserted. But these studies inspire to seek solutions to once again captivate readers in the literary world.

1.3 Motivation

It is clear that the situation requires changes and shock plans to improve the process of selling books to readers, but with those who really captivate with their reading. This constitutes my main motivation for this research work. Focusing on the idea of creating a recommendation system that focuses on the personalization of the books that are recommended, benefiting those involved in this process (editors and readers), but especially the publisher. Not only would this improve publishers' profits and the buying process (by reducing the time and cost of finding the desired book); it will also stimulate the reader's empathy with the book. Readers could become the character of the book itself. And consequently, it would reinforce self-esteem and confidence in one's own person. There is nothing like recognizing yourself between the pages of a book to feel special.

1.4 Problem Statement

The situation described above is the justification that responds to the question posed by the nature of the problem: Would designing a book recommender based on readers behavior encourage more reading?

In general terms, to design the recommender, certain points must be taken into account, such as:

1. **Literary Genre:** Since Non-fiction represents a specific aspect of the real world, the factors that influence its publication (hence, sales) are more closely associated with global characteristics of the book itself. However, in the case of fiction books, since it is an imaginary world inhabited by mostly invented characters, several factors associated with the story itself have an influence.

2. **Language learning:** The vast majority of the bibliography (as can be seen in the Chapter 2) is based on the English language, and although there are some technological advances in the rest of the languages (specifically in Spanish) adjustments are still required for intelligent systems to be able to understand certain inflections that are a product of the language itself.

3. **Sentiment analysis:** At this point, it not only refers to the classic classification of a positive or negative, but also to open up a little more the range of possibilities that natural language processing technologies offer us today; and to be able to determine other psychological factors, both at the level of the reader and the characters of a literary work.

4. **Reader behavior:** Based on user behavior on social networks, since it is where society currently develops the most. Then, analyzing its form of expression in these media, since it is variable depending on the period of time in which it is found.

1.5 Hypothesis and starting point

In order to create a recommendation system that is more adjusted to the needs of the readers, it is essential to improve the feedback process. It is vitally important to collect information from readers; and find analogous characteristics between the characters of the work and the reader. Based on the arguments described, the following is proposed as a starting hypotheses:

It is possible to build a book recommendation system that adjusts to the needs of the readers and allows them to enrich their reading and satisfaction.

It is possible to implement a recommendation system according the differential mode of fiction and non-fiction books.

1.6 Objectives

Based on the proposed research framework, this section will clearly and concisely detail the result that is expected to be achieved with this thesis; as well as the more concrete and measurable objectives that are expected to be achieved with this thesis in a specific period of time and always at the service of the main objective.

1.6.1 General objective

Therefore, the general objective that is derived from the antecedents is:

Design a book recommender based on readers behavior that encourages greater reading and satisfaction.

1.6.2 Specific objectives

To achieve our general objective, the following specific objectives are set:

1. **Segmenting the recommendation according to the literary genre:** As we have explained before, the book recommendation approach based on literary genres turns out to be quite different from each other. On the one hand, the Non-fiction genre is based on the characteristics of the book: price, theme, etc. and on the author's behavior on social networks. And on the other hand, the Fiction genre is based on the story that is narrated.

2. **Influence of an author in social networks:** Since it allows real knowledge of the interaction between an author and readers and/or followers; and with it, analyze book sales. Even extend it to the mentions of said author on the Web, since an event can generate a high spike in sales.

3. **Create a character profile of a work:** With this, the system will be able to establish a relationship between the characteristics of the readers with that of the characters of the work, and therefore capture the reader's attention.

4. **Design a model capable of not losing any information about the characters:** the design and development of a model that does not lose any characteristics of the characters, due to the result of diminishing the cacophony in the work.

5. **Develop a recommendation platform:** This provides support for online sales platforms, recommending books based on readers behavior.

1.7 Document structure

The thesis is organized as follows. In Chapter 2, useful concepts are presented to understand the methods applied in future chapters, such as classification, regression in those related to sales predictions; an introduction to the world of natural language processing (NLP) and psychology. The current situation of obtaining a psychological profile of both characters and readers and the basis for understanding recommendation systems is also reviewed. In addition, a review of the works related to all the topics addressed in this research work is carried out. As one of the specific objectives is the recommendation of books according to the literary genre, in Chapter 3 we will analyze the Non-fiction genre. During this chapter the first contribution made during the development of this thesis is exposed. Both the characteristics of the book are analyzed, as well as new indicators are added, such as the author's behavior in social networks. In Chapter 4, the results of the second contribution related to the psychological profile of a Fiction book character are presented and analyzed, which will lay the final basis for the adjusted book recommendation system to the needs of the reader that is developed in Chapter 5, this chapter shows the result of the experiment carried out. The assessment of each of the contributions made can be seen in Chapter 6. Finally, in Chapter 7 the general conclusions and the main future lines of work derived from this research are presented.

Chapter 2

Theoretical Background

"It all depends on how we see things as they really are."
-Carl Gustav Jung-

This chapter presents the conceptual bases for the understanding of the whole thesis; plus the current and future situation of book recommendations: the main indicators for the prediction of book sales are distinguished; it is studied in the semantic architecture of texts in Spanish; the psychological profile of each of the characters in the works is analyzed. In addition, the bases for creating a recommendation system based on experience cases are studied.

2.1 Introduction

During this chapter, the basic concepts are exposed to understand how the system developed in this thesis or research project operates. Also, a review of the literature is carried out, in order to show that it currently exists in the academic framework and in the current market. Likewise, it provides the bases to demonstrate the novelty of the results of this research, which are: prediction of book sales; the characterization of the personality of the character of the work and the recommendation of books for the purposes of the interests of the readers.

2.1.1 Machine learning

The ability of a computer system to learn automatically is essential in Artificial Intelligence (AI), since for a system to be considered intelligent it must be at least capable of learning automatically [7], hence the term Machine learning (ML), which is a branch of AI, see Figure 2.1.

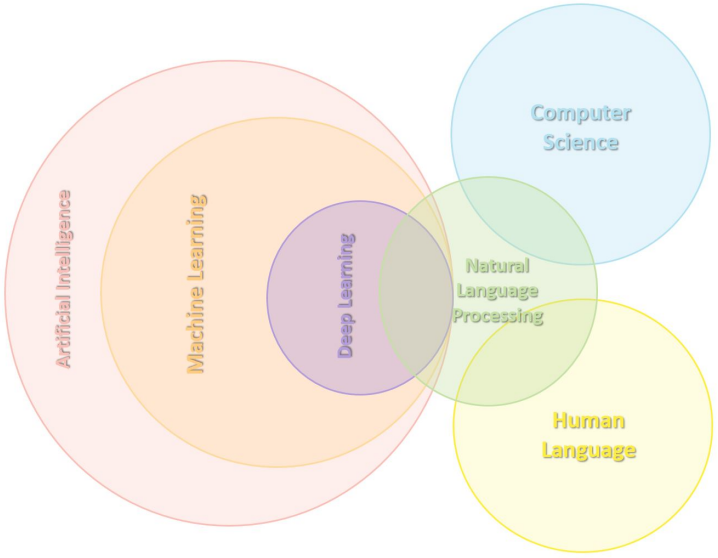

FIGURE 2.1: Interrelation of the branch of Artificial Intelligence and Natural Language Processing. Source: Own elaboration.

The techniques and algorithms used in the modeling of ML applications are based on theories of statistics and probability, thus taking advantage of the computational capacity of current processors to execute calculations iteratively and in parallel on large amounts of data. On the one hand, statistics uses mathematical methods to discover relationships between data points while ML relies on statistical methods to design a model that can predict future events, classify existing information, or detect patterns among data.

These algorithms are classified into the following categories:

- Supervised learning algorithms: are those that make predictions based on previously labeled training data. Each training dataset sample includes one input and one output. Allowing the algorithm to parse the data and make an inference by determining the unseen data labels. These algorithms can be subdivided into two types: classification (qualitative values) and regression (quantitative values), which will be used in Chapter 3. [8]
- Unsupervised learning algorithms: are those that discover knowledge and relationships in unlabeled data. These models receive input data, but the desired outputs are unknown; so they have to make inferences based on circumstantial evidence, without any guidance or training. Using for this task such as clustering in order to achieve the grouping of sets of unlabeled objects (k groups). Models are not trained on the "correct answer", so they must find patterns on their own. One of the most common algorithms is KMeans, which will be used in Chapter 3. [9]
- Semi-supervised learning algorithms: are those in which the training data is divided into two. A small amount of labeled data and a larger set of unlabeled data. [10]

2.1.2 Deep learning

Deep Learning or deep learning is a subset of the family of methods of ML that are themselves a subset of the wider field of the AI, as observed in Figure 2.1. It is based on artificial neural networks (ANN), a type of computer system that emulates the way the human brain works. When a model receives input data, which can be images, text, video, or audio, and is asked to perform a task, the data is passed through multiple layers of interconnected neurons, allowing the model to progressively learn as it does a human does by evolving with experience. These neural networks will be used in Chapter 4 along with Natural Language Processing (NLP). [11]. For a better understanding of the deep learning concept, see Figure 2.2.

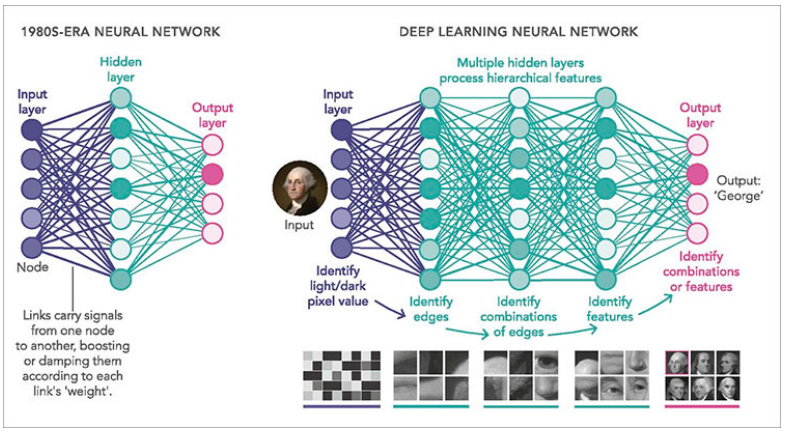

FIGURE 2.2: Difference between simple and deep neural networks. Source: https://www.futurespace.es/en/redes-neuronales-y-deep-learning-capitulo-1-preludio/.

Among the many fields in which deep learning can currently be used are:

- voice recognition
- computer vision
- audio recognition
- natural language processing
- social media filtering
- machine translation
- bioinformatics

In some cases comparable or even better results have been achieved than the human experts.

2.1.3 Natural Language Processing

Natural Language Processing (NLP) accords to Keselj [12] is a subfield of linguistics, computer science, and AI that deals with the interactions between computers and human language, in particular, how to program computers to process and analyze large amounts of natural language data. The goal is to have a computer capable of "understanding" the content of documents, including the contextual nuances of the language within them. Named Entity Recognition (NER) is popularly used in NLP tasks, which provide much-needed data classification and interpretation assistance.

As it has been observed in Figure 2.1, NLP was born as a subarea of AI and Linguistics. It has evolved over the years. Below is a brief and chronological description of each of the algorithms developed for this field:

- 1954 BoW: In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but maintaining multiplicity. [13]
- 1954 TF-IDF: In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but maintaining multiplicity. [13]
- 1974 : It is a numerical statistic that aims to reflect how important a word is for a document in a collection or corpus. [14]
- 2013 Word2vec: The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. [15]
- 2014 GloVe: It is an unsupervised learning algorithm to obtain vector representations of words. In training, the global aggregate word-word co-occurrence statistics of a corpus show interesting linear substructures of the vector space of words. [16]
- 2016 Fast Text: The model allows you to create an unsupervised learning or supervised learning algorithm to obtain vector representations of words. Facebook makes available pre-trained templates for 294 languages.[17]
- 2017 Transformers: It is a deep learning model that adopts the self-attention mechanism, differentially weighting the meaning of each part of the input data. [18]
- 2018 ELMo: Represent a sequence of words as a corresponding sequence of vectors. The tokens go into an LSTM. Embeds are context sensitive and produce different representations for words that share the same spelling but have different ones. [19]
- 2019 BERT: It is a transformer-based machine learning technique.[20]

2.1.4 Recommender system

Recommender systems are important tools that help users learn about options or items of interest to personalize the user experience. Every day, people come into contact with these powerful recommender systems [21]. When you enjoy a video on YouTube or let Spotify do a mix of artists for a playlist, you are providing customization elements for these systems to build their recommendations.

Recommender systems combine ideas of information retrieval and filtering, user modeling, machine learning, and human-computer interaction. The following will briefly explain what the two main classes of recommender systems consist of:

- Collaborative Filtering (CF) : This system generates recommendations using only information about rating profiles for different users. Collaborative systems locate peer users with a similar rating history as the current user and generate recommendations using this neighborhood [22–26]. For a better understanding, Figure 2.3 shows an example of the operations performed by a recommender with collaborative filtering.

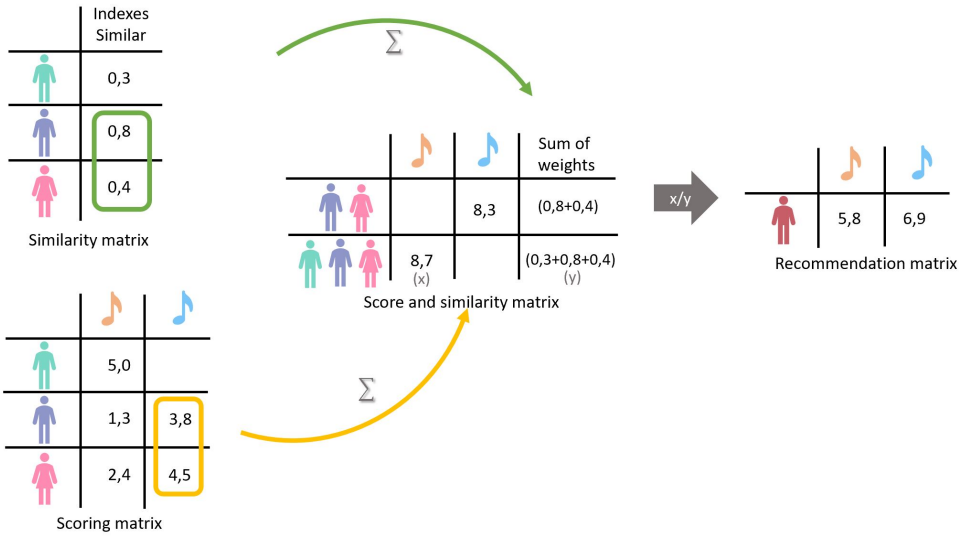

FIGURE 2.3: Operation of a recommender based on collaborative filtering. Source: Own elaboration.

In the figure above, it can be seen that in order to recommend the songs to a new user (red user), a score and similarity matrix must have been previously created with respect to other users. The resulting values will be saved in the similarity matrix after calculating the distance between these users (purple, pink and blue user) and the new one (red user).

The similarity metrics commonly used and the selection of each one will be given by the following conditions: [27]

  - Pearson: when data is subject to user bias/different user rating scales.
  - Cosine: If the data are scarce (many classifications are not defined).
  - Euclidean: If the data are not scarce and the magnitude of the attribute values is significant.

And finally, in the scoring matrix, there is the vote given by the users to each of the songs.

To create this matrix, once the users (purple and pink users) who voted for the blue song are known, their votes are added. And in addition, a sum of the weights corresponding with these users (which are found in the similarity matrix) is calculated. Having these values, it is divided between them, and it will show the optimal recommendation for the new user.

- Case-Based Reasoning(CBR) : A recommender system is initially based on case-based reasoning (CBR), which consists of solving new problems based on the solutions of previous problems [28–30]. It could be simplified by commenting that CBR is a way of reasoning by making analogies. CBR is fundamentally based on four phases: Retrieve; Reuse; Review; Retain the continuity or follow-up order of each of them. Figure 2.4 shows each of these operations.

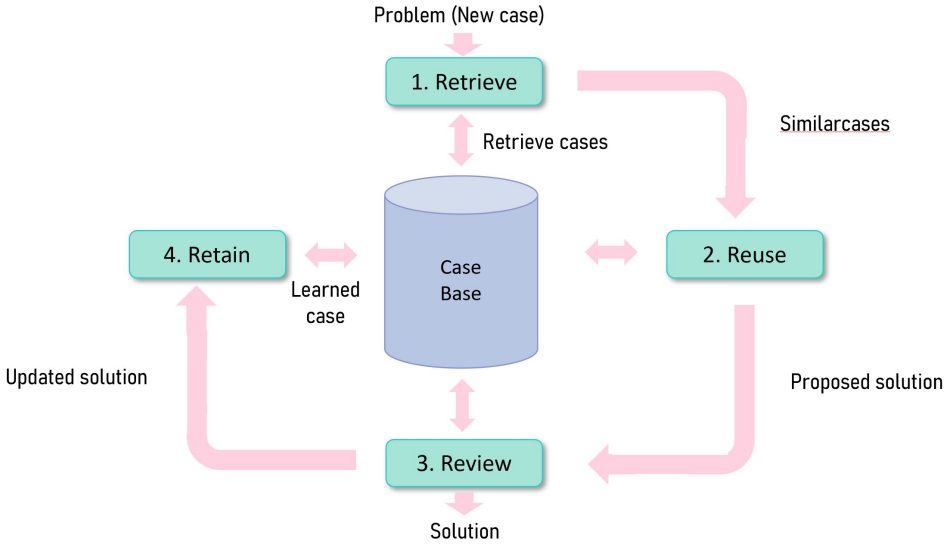

FIGURE 2.4: Operation of a recommender based on cases. Source: Own elaboration.

The CBR technique could be summarized in two main blocks according to its functionality: a classifier and a synthesizer. One of the classical advantages of CBR is the simplicity of its classifier, with a K-Nearest Neighbors (KNN) [31, 32] classifier being a common choice.

Explaining what the two classes of recommendation systems consist of, in the investigative work that will be developed, the CBR is chosen since it will be based on the hidden similarities between the readers and the books; and not in the ratings/qualifications that can be given to the books.

2.1.5 Other concepts

Psychology is the science that studies mental processes. The word comes from the Greek: psycho- *(soul or mental activity)* and -logy *(study)*. This discipline analyzes the three dimensions of the aforementioned processes: cognitive, affective and behavioral. There are two instruments to assess personality (which come from written questions that a psychologist asks the participant about the three previously mentioned dimensions):

- Myers-Briggs Type Indicator (MBTI): Is a pioneering work by Katharine Cook Briggs and Isabel Briggs Myers [33] in the field of psychological typing. It is a very popular assessment used to determine personality in the dominant culture.
- The big five (B5): Is a taxonomy or classification of personality traits that analyzes the composition of five personality dimensions in their broadest sense. These factors were found experimentally by Golbert [34] in an investigation of the personality descriptions that some people made of others.

So what is the difference between these two measurement methods?

Self-assessment (MBTI) focuses on how an individual measures their own internal dynamics. On the other hand, observer ratings (B5) focus on how a peer or neutral observer perceives an individual's personality. Therefore, observer ratings are based more on social behavior, meaning that it is determined by how an individual is perceived by others; whereas, self-assessments are shaped by one's own feelings and motives.

2.2 State of the Art

The current panorama of technologies based on the three areas to be dealt with during this thesis is detailed below.

2.2.1 Sales predictions

The concepts acquired in the previous section will allow the understanding of the entire project and the use of this branch of AI in the publishing sector.

Current situation

The Influence of Social Networks on Sales

From the perspective of social influence, no study of the publishing sector considers the impact of social networks on the product. Still, there are many similar cases in different retail industries, for example: the fashion study presented by Park et al. [35], shows that a strong social media presence is important for selling a product, rather than being under industry advertising standards. Another case would be in mobile telephony, there is a study by Lassen et al. [36], indicating how these communication channels can be used to predict the revenue of a product. And lastly, in entertainment, as found by another study, this time by Abel et al. [37], where the authors reveal that beyond social networks, information can be extracted from celebrity blogs. Specifically, the paper highlighted by Moon et al. [38] proposes to use the number of blog references as an indicator of the success of a book's sale. Rapp et al., in their study [39] understands the effects of social media on seller-consumer interactions, offering new insights into how social influence enhances consumption impacts and positively contributes to retailer performance and consumer loyalty. Following this line of thought is the Guesalaga article [40] based on customer engagement with a product or brand, which reinforces the importance that this feature may have for future predictions. Recent publications such as the one by Barabasi et al. [41] serve as a foundation to start to understand the mechanics of reader preference. Even so, they are only based on the sale of bestsellers, thus leaving a large gap in the sale of the remaining books. Another article like the one presented by Kim et al. [42] makes us reflect on the effect that sharing the famous "likes" of consumers on social networks has on sales, exerting great social pressure in the community.

The success of a book based on sales

From the point of view of the success of the book based on sales, the work carried out by Barabasi et al. [43] analyzes the characteristics that make a book a bestseller, using statistical and data analysis techniques. But the work is aimed exclusively at authors already recognized by readers. Another study carried out by Feng et al. [44] is only based on historical book sales data and some attributes collected from Amazon. At this point in the research, none of the sources include the publishing industry's social media to predict sales. However, they help us to consider what features should be introduced into the prediction model.

Sales segmentation forecasting techniques

The goal of segmentation is to divide heterogeneous groups into homogeneous subgroups based on similarities. One of the most used statistical techniques for this purpose are the quartiles proposed by Rew [45], which divide populations into four groups, or quarters, of more or less the same size. All the work proposed by Winfrey et al. [46] allows analysis using quartiles for market segmentation, but it is only focused on the buyer. The consumer market in this paper is segmented by a price, showing that wealthy buyers pay moderately higher prices for pills and injectables. Another study by Lehmann et al. [47] reveals a negative wage gap in the lowest quartile of the wage distribution of a labor balance in Russia. These articles are useful as they provide us with a segmentation technique that we will use to compare with the expert segmentation (currently used) in the publishing industry. Finally, although following this line of research, the works [48–50] where pattern matching techniques are applied to time series data are interesting, because they use the similarity of historical data as a basis for pool time series. The previously described makes it possible to visualize an idea of grouping historical information from books, but we still need to incorporate data from social networks and the web.

2.2.2 The personality of the characters

In this area to be dealt with, it is of vital importance to take into account the concepts of NLP and Psychology previously commented.

Current situation

NER at present

The detection of characters in literature is a constant improvement, in which little by little progress is being made in the field of investigation. This was especially due to the presentation in 2017 by Vaswani et al. [51] of the concept of attention, which allows working in parallel with each of the words, without having to lose its essence as the text is read. Later, in 2019, the bidirectional encoding representation of transformers (BERT) appeared by the Google development team. Vala et al. [52] present an approach to the use of this model, where they explain how, through graphics, it is possible to determine the degree of association or compatibility that the characters of a work have, where different heuristics are applied for this achievement. This relationship is quite subjective, due to the problem that arises since they do not require the information that each character presents, but they help to take into account certain basic considerations for the model that will be created. On the other hand, Wei et al. [53] are based on the search for personal pronouns using disambiguation rules, although 100% are not taken into account since the language they work with is Chinese; and we know that this language does not have the same Subject-Verb-Object (SVO) patterns as Spanish. But the study allows thinking about the use of maximum entropy for the identification of entities. Furthermore, the work of Chen et al. [54] shows another way to approach character recognition in a text, instead of statistical models, they use rule-based models. Again, it does not solve the research question that was initially raised in this investigation, since in this case dialogues from television programs such as The Big Bang Theory and Friends are used and these in the end present a totally different structure than the one presented in a fiction book.

Personality profiles of fictional characters

The emotional contagion that a book evokes is an important feature of readers' emotional engagement, as the study by Tirre et al. [55], and not only adds to the literary work, but helps the writer create a more realistic character. That is why it is of vital importance that experts in the publishing sector create a profile of a character. For example, Flekova et al. [56] uses the focus group as a method to determine the personality of the fictional book's protagonist, also based on the idea described by Propp. While this is a method that will not be used in this study, it provides the foundation for understanding how personality can be linked to character and what traits might be important, such as a character's actions. Furthermore, the study by Elson et al. [57] guarantees that characters can have relationships with each other if they are within a range of 300 words. This allows within the profile of a character to recreate relationships with others. The personality of a character can be seen reflected in the studies [58–60], which achieve a 50% accuracy in text and 45% in speech, of the determination of five personality traits, to create a basic starting point, but without taking into account the characteristics that can be found in the personal pronouns used by writers to avoid cacophony.

Equivalences or contrast between the most spoken languages English, Chinese and Spanish

At first glance, the sign grams presented by the Chinese language are not similar to Spanish or English. But beyond the strongest evidence, the grammar is not analogous either, since Chinese puts the object before the verb. Figure 2.5 shows an example of the grammar used in Chinese.

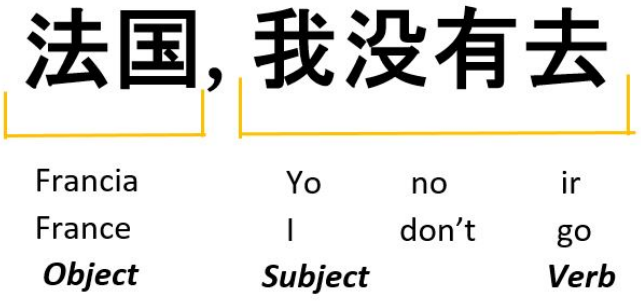

FIGURE 2.5: Differences in the grammatical order between the Chinese, English and Spanish languages. Source: Own elaboration.

Fortunately, both English and Spanish follow the same Subject-Verb-Object (SVO) grammatical pattern, but this does not mean that differences still exist when comparing these languages. Cantos [61] proposes an in-depth analysis of this topic, where it points out that one of the differences lies in the inflections. For example, unlike Spanish, in English gender inflections are not observed because this language offers a more simplified table of forms, but you can see the amount of inflections relative to the number. Another divergence that these languages present is found in the length of the words/sentences. And, last but not least, there is the notable difference in lexical richness (which would be equivalent to a wide communication capacity), although it can be treated in a relative way since it will depend explicitly on the societies that speak them. For example, the term "paella" does not exist in English, since it is a typical dish of Spanish speakers. It is shown that the language with the highest number of lexical words is the English language, since a single word can have dissimilar meanings. Example of this,

- *He gave her a ring with a huge diamond!* (in this case, ring, refers to the noun ring)
- *Give me a ring later, we'll talk about it.* (in this case, ring refers to the verb call)
- *Ring the bell if you want them to open the door for you.* (in this case, it refers to the verb to spund of a bell)

In addition to the similarities and differences between these languages, this work allows understanding the importance of properly designing a linguistic corpus according to what is going to be investigated.

2.2.3 Recommendation systems

In a global way, and as reflected in the specific objectives, the general idea of this research project is to create a book recommender system (for both the Non-Fiction and Fiction genres). That is why it is important to see the current trend and existing technologies in this field.

Current situation

CBR origin and evolution

The first contributions in the area of case-based reasoning (CBR) were from Schank et al. [62], where the use of CBR as a conceivable high-level model for cognitive processing was highly appreciated. Later, in 1994, Aamodt et al. [63] successfully used it in several domains, such as prediction, control, and planning. With the evolution of AI, many works focus on incorporating this type of reasoning into AI systems: [64–68] among other studies that have been developed to date present. Recently, a study by Adelomou et al. [69] has linked case-based reasoning with the quantum world, where more complex algorithms combined with a quantum are used to determine similar cases in the Recovery phase. If the review focuses on the publishing sector, the study carried out by Chang et al. [70] even offers us the possibility of merging Artificial Intelligence algorithms within the same system to improve its accuracy. Research by Chen et al. [71] shows that customers are more interested in books labeled "customers who bought this book have also bought" than books labeled "recommended by bookstore staff". In this case, it combines it with unsupervised machine learning techniques such as self-organizing maps (SOMs) to forecast book returns to publishers. Later Chang et al. [72] offers the same approach, this time achieving a hybrid between genetic algorithms (GA) and CBR. From the literature reviewed, it is observed that this type of techniques is not sufficiently exploited in the publishing field, especially in the recommendation of books that meet the needs of the reader, based on the writing, since the few that exist usually deal with this type of problem with a global aspect such as the most frequent sales and purchases.

Combination of techniques: CBR using NLP

As the work focuses on recommendation systems in the publishing field, specifically based on the text of books, the literature is searched for studies referring to the combination of CBR techniques using NLP. Only one relevant study has been found, although it was outside the editorial domain. At least, the authors of this research have not found related works to date. The first study by Wu et al. [73], allows the recovery of metro accident cases through the development of an ontological model using the NLP technique for decision-making, given that the memory cases consisted of historical records of accidents. This allows, although not from the same domain, to get a little closer to the idea of this research and delve into the similarity of the texts to provide similar cases of a CBR.

Similarity of texts in Spanish through NLP

For the recommendation of books, based on the writing, it is important to analyze the text both at the lexical level and at the semantic level. The use of one or the other will depend on the task to be carried out, for example at the lexical level the words will be organized in groups or fields of meaning and at the semantic level they will be associated because they belong to the same grammatical category and share a part of their meaning. As an example, Figure 2.6 shows the difference of both levels.

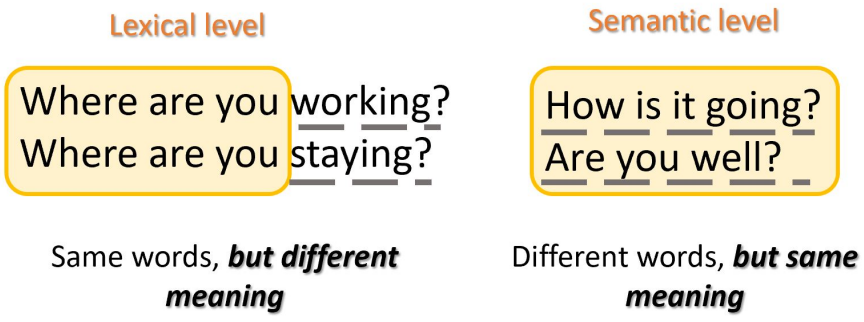

FIGURE 2.6: Lexical analysis versus semantic analysis in texts. Lexical analysis is based on similar words even though the meaning is different. While the semantic analysis, although it has different words, they present the same meaning. Source: Own elaboration.

However, the vast majority of studies, such as [74, 75] on text similarity have focused mainly on the English language. It, which means that there are not as many resources as would be needed when working with the Spanish language. Of the few studies found, there is a work of Lopez et al. [76], where they use the vector representation based on word embeddings for the task of semantic similarity of texts, using evaluation metrics such as Euclidean distance and cosine. Finally,

with all these studies, a base of the evaluation metrics and the different approaches that can be taken when analyzing a text are taken, and the limited resources that still exist in the Spanish language are reaffirmed (despite being the 4th most spoken language worldwide) as can be seen in Figure 2.7. The use of this similarity in the Spanish language, as will be seen in Chapter 5, will allow the Recovery Phase of the CBR to find books that resemble the behavior (expressions) of the readers.

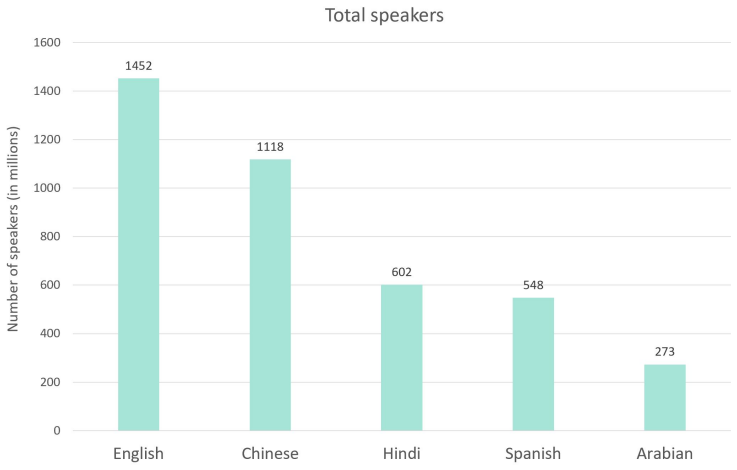

FIGURE 2.7: The most spoken languages in the world, with Spanish being the 4th most spoken. Source: Own elaboration based on https://es.statista.com/estadisticas/635631/los-idiomas-mas-hablados-en-el-mundo/

Zero shot learning

Finally, it is searched within the literature how to recommend both Fiction and Non-fiction books. Since the latter is closely related to the topics currently being discussed, it was necessary to understand how the texts could be labeled in an unsupervised way. That is why the concept of zero-shot learning is studied, which is a configuration of problems in machine learning, where at the time of the test, a student observes samples of classes that were not observed during training and needs to predict the class to which it belongs. Reviewing the literature, a study by Yin et al. [77], which provides the basis for understanding this technique, offering as its own contribution a dataset created to facilitate the study of this technique. It proposes global and then disaggregated tags, where you can see how challenging this technique can be if used with very specific tags. Subsequently, a study by Pan et al. [78] that presented promising results thanks to the availability of large datasets, but again there is the linguistic component that influences the application of this type of technique since they are specifically modeled for the Anglo-Saxon language. Finally, there is a multilanguage model [79] that presents the best results so far since it allows extending the training and test datasets of the Corpus Multigenre Natural Language Inference (MultiNLI) to 15 languages, allowing this

to be used for the Spanish language, which is the one of interest during this project, and which will be used to show the operation of book recommendations based on the topics covered.

From the literature reviewed in this chapter, you can: identify the latest advances in research and technology and discover that the integration of two or more algorithms could increase the forecast accuracy effectively. Therefore, it is what is developed in the following chapters.

2.3 Summary

In this section, the reader of this thesis has been placed in the current context in which technology is found, in the three contributions that will be addressed in this project: Non-Fiction Books (NFB), Fiction Books (FB) and CBR. (Case Based Reasoning), as can be seen in Figure 2.8.

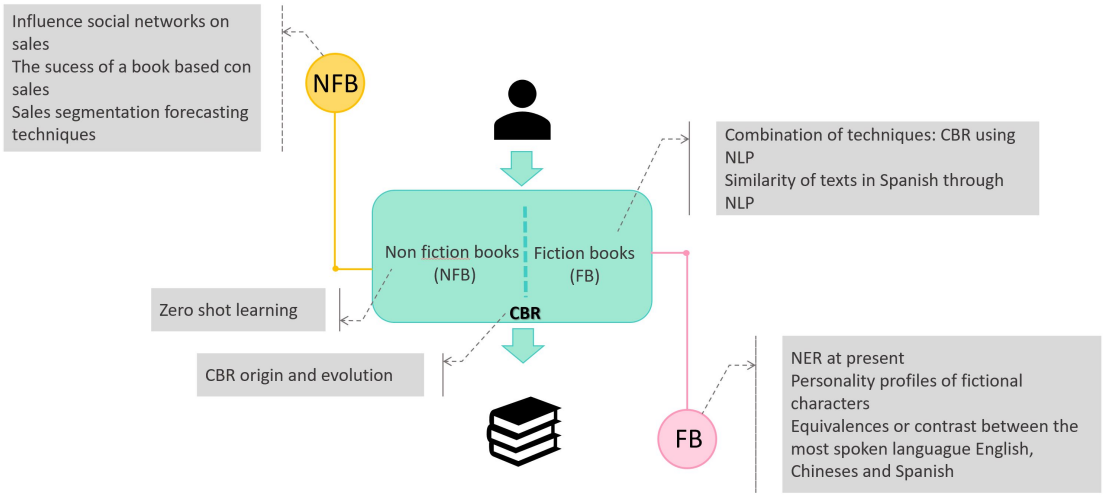

FIGURE 2.8: Techniques of the state of the art in the final project. Source: Own elaboration.

They are all intertwined with each other, as the project includes recommending both Fiction and Non-Fiction books. Because these genres are analyzed from two different perspectives: one at the level of interaction and interests of the reader in today's world and another based on the content (characters, setting, etc.) of the work.

Part II

Research work

Chapter 3

Sales Prediction

"There is no sale without the story; no knockout without the setup."
-Gary Vaynerchuk-

This chapter presents and details the first contribution made with this research. It consists of the creation of a combined model of Artificial Intelligence techniques for the optimal prediction of copies of the Non-Fiction genre to be printed. This prediction is based mainly on the behavior of stakeholders in social networks and the web; which allows the predictions to be adjusted much more compared to the current method under the criteria of the experts.

3.1 Introduction

Non-fiction is a literary genre characterized by the use of exposition, description, narration or argumentation of a true content or based on real events. The same happens with the Children's/Youth literary genre, since this genre uses different resources to express reality and educate minors. This is why the sales analysis is carried out for these two literary genres. Therefore, it is closely linked to today's society, specifically, with the new technologies promoted by the Internet, such as social networks. During this chapter, these factors are experimented with as indicators of the book's success.

3.2 Dataset

The data used for the experimentation is provided by (what we will call) "The Editorial", respecting the anonymity of the medium; and correspond to Growth From Knowledge (GFK) data on book sales based on the behavior of authors on social networks. The definition of each of the characteristics belonging to the book is described in Table 3.1. On an initial scan of the data (especially of the book features), it can be seen that the dimensionality of the model would be too high. Therefore, to identify and eliminate irrelevant variables, a variable correlation analysis is performed. To do this, initially the nominal categorical variables are transformed into numerical variables to be used later in the statistical analysis of book sales. This transformation is done through techniques such as Label Encoding, which allows converting the labels in a numerical format to convert them into a machine-readable format.

For the analysis of correlation of variables, there are two common measures: Spearman's correlation and Pearson's correlation. It is decided to use the latter, since to use Spearman's correlation it is required that at least the variables are in an ordinal scale. The Pearson [80] correlation is a statistical procedure for determining whether two variables are linearly related. And its formula is defined as Equation 3.1.

$$\rho_{X,Y} = \frac{\delta_{XY}}{\delta_X \delta_Y} = \frac{Cov(XY)}{\sqrt{Var(X)Var(Y)}} \tag{3.1}$$

where $\rho_{X,Y}$, is Pearson's population correlation coefficient; δ_{XY}, is the covariance of XY; δ_X, is the standard deviation of X; δ_Y, is the standard deviation of Y.

TABLE 3.1: Description of the basic characteristics of the GFK dataset book.

Features	**Description**
ean/isbn	A type of 13-digit barcode used internationally to identify each book.
genre1	Correponding to the general literary genre of books. *Example: Fiction, Non-fiction, etc.*
genre2	Correponding to the literary genre in the second level of the books. *Example: Lifestyle, Human Sciences, etc.*
genre3	Corresponding to the theme that the book deals with within the literary genre. *Example: Self-help, Gastronomy and Cooking, etc.*
format	Format of sale of the book. Example: Hardcover, Softcover, etc.
region	Region in which the books are sold. Separated by autonomous community.
channel	Book sales channel. *Example: Amazon, Bookstores, etc.*
week/year	Week and Year of book release.
units	Units sold of book.
value	Earnings of units sold.
unitValue	Price at which the book is sold.

The result obtained is a correlation coefficient with values that are between -1 and 1. The closer to 1 they are, the greater the positive correlation will be (the magnitudes increase in the same way) and the closer to - 1 greater will be the negative correlation (the magnitudes of the first variable increase and those of the second decrease). In order to carry out a good correlation analysis, a normalization of the data is carried out. Specifically, a min max normalization, which scales all data in ranges of [0,1]. In this way, the study of the importance or correlation of the variables is real and there are no biases due to data scales higher than others.

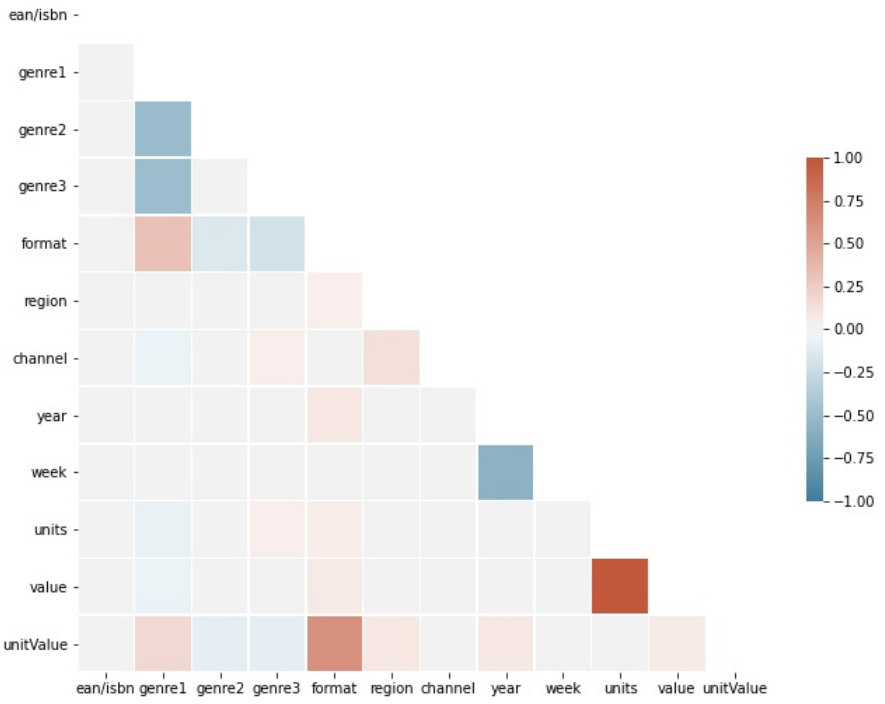

FIGURE 3.1: Correlation matrix of all the variables belonging to the characteristics of the book (GFK). Source: Own elaboration.

The results in Figure 3.1 show that there is a strong connection between sales (units) and price (`value`). Despite the lack of a sufficiently significant correlation with the rest of the variables, it is possible to intuit and draw conclusions about which ones will be relevant for this research. Although it is true, and this is confirmed by the data shown, sales have a very close relationship with prices, but they also have a slight relationship with the type of theme (`genre3`) they offer. If the unit sales price itself is analyzed, it is observed that it has a significant relationship with the format and the year of publication (`yearRelease`). There is no doubt that there is no correlation between sales or price with the variable (`weekRelease`), but it was originally linked to the variable (`yearRelease`), see Table 3.1. Thus, it allows authors to determine that the publication date (week and year) is an important factor in sales. Furthermore, new variables are created from this factor, such as (`numberBookPublished` and `weeksBetweenPublication`). In addition, implicit information such as (`title` and `author`) is added to this selection of characteristics, which do not appear in the indications because they are categorical variables.

However, to analyze the correlation of the variables during the data preprocessing, a Machine Learning model is applied: XGBoost Regressor; for the final selection of the variables to use. This model was selected because by using gradient enhancement and once the enhanced trees have been built, it is relatively easy to retrieve the importance scores for each attribute. What finally allows the comparison between the attributes and select the optimal for further experimentation. The results after applying this model can be seen in Figure 3.2

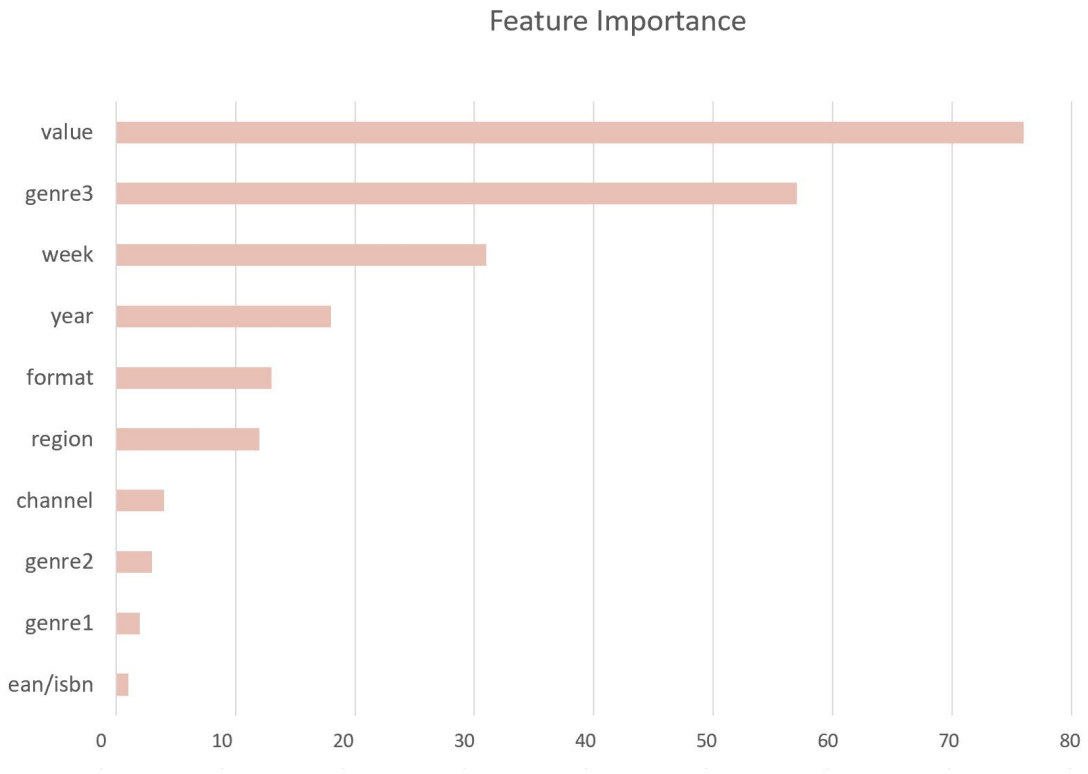

FIGURE 3.2: Importance of the variables belonging to the characteristics of the book (GFK), through the application of XGBoost Regressor. Selecting the first 4 for having an importance greater than 15%. Source: Own elaboration.

It can clearly be seen that both criteria coincide, both the correlation analysis and the selection from the application of the Machine Learning XGBoost model. The main numerical characteristics are selected: the price, the genre, the week and the year of release.

Naturally, book sales are closely linked to the public interest, so it is essential to include this variable in the sales analysis. According to the review of the literature carried out, with social networks, it is possible to retain customers in a much more effective way than with other media and, in addition, it will be just as effective in smaller social networks, such as some created ones on certain hobbies, where it will be easier to establish a direct channel of information with the client. Through the interaction of authors with their audience on social networks, they can have a greater impact with their work. An example of this is the Self-help books (Non-fiction genre) by author Marian Rojas, in Figure 3.3 it can be seen the number of followers of this author.

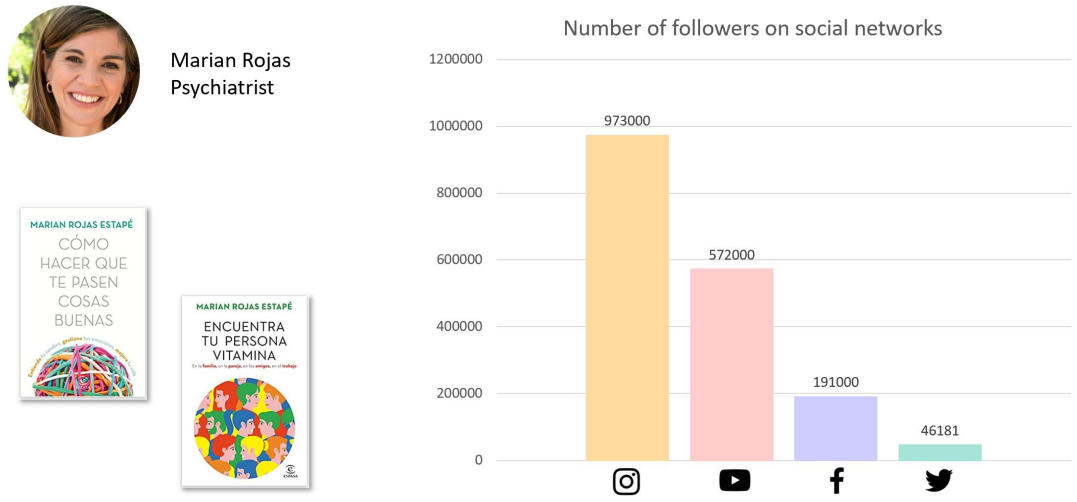

FIGURE 3.3: Distribution of the number of followers of Marian Rojas on the different social platforms. Reaching higher volume on Instagram. Source: Own elaboration, adapted from: https://album.mediaset.es/eimg/10000/2021/09/15/clipping_S9y8bc_15aa.jpg?w=1200&h=900; https://m.media-amazon.com/images/I/6177q+2bb0L.jpg and https://m.media-amazon.com/images/I/712cARV+H7S.jpg

Likewise, the phenomenon of mentions on the Web is included as another factor to be analyzed in sales. As many will know, in the field of culture, a painter, writer, among others, they become famous after his death, or with an award ceremony or even an interview, event or news about the author and his or her work. Continuing with the author of the previous example, Figure 3.4 shows the trend of the author's mentions during the period 2018-2022 in Google. Where the peaks marked in green are a product of the marketing carried out on the launch dates of each of her books; the peaks marked in blue are the consequence of the holiday and gift campaigns for Christmas and/or Epiphany; the peak marked in red, a result of marketing related to Valentine's gifts; the peak marked in purple is the result of marketing during the Book Day dates and finally the orange peak is the effect of her presentation at CIC2021.

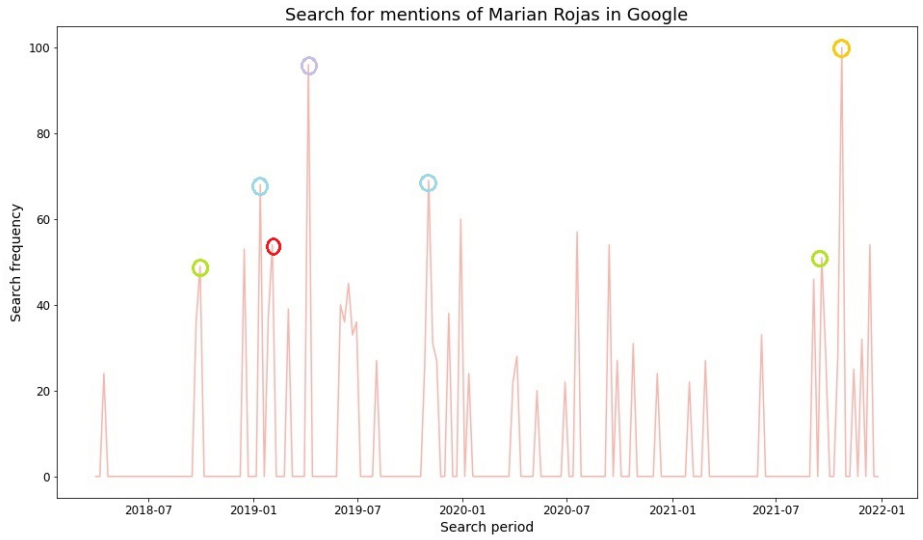

FIGURE 3.4: Search trend of mentions of Marian Rojas in Google. Source: Own elaboration.

If the impact of these mentions on the Web is analyzed, following the same example with the author Marian Rojas. It can be seen in Figure 3.5, that during the months of October-November there is a rise in sales as a result of the launch of her first book. Later, there is another spike during Valentine's Day, which is very valid since it is a period where many couples seek to give a book as a gift. And then if we look at the resounding growth during the Book Fair dates, it is observed that it coincides significantly with the previous searches. Finally, the impact of Christmas marketing during the month of December 2019 can also be visualized.

FIGURE 3.5: Marian Rojas book sales distribution. Source: Own elaboration.

The previous examples demonstrate the importance of including these metrics within the characteristics of the books for the predictive analysis of their sales. Finally, this dataset is complemented with information on the behavior of the authors in social networks and their mentions on the Web. The data are limited to the territory of Spain during the period 2018-2020 and to the literary genre Non-fiction. As a sample, 1169 books have been used, which according to experts in the publishing industry, are considered more sensitive to the popularity of the author. The analysis is based on the characteristics of a book, the author's public social networks and web mentions, as seen in Figure 3.6.

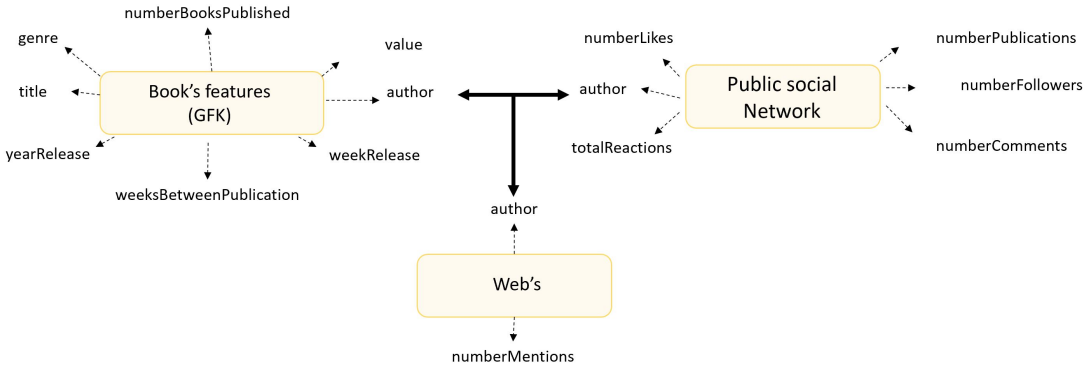

FIGURE 3.6: The entity-relationship diagram of the various sources works to form the input dataset to the proposed system. The bidirectional arrows indicate that the author is the key to the union between the different sources. Source: Own elaboration.

3.3 Proposed tool

With the previous dataset, it is proposed to compare different segmentations to determine which is the optimal for copy prediction. For a better understanding of what is proposed with the experimentation of the study [81], Figure 3.7 shows how, once the data is entered into the CAIT system, they will be classified into the different segmentations and individual regressors will be applied to each one of these. The defined segmentations are: (1) the quartiles (influenced from the perspective of the experts); (2) segmentation by the expert (there are 4); and (3) segmentation by grouping using a clustering technique. Coincidentally, in the last segmentation, the k groups give a value of 4, as will be observed in the experimentation section.

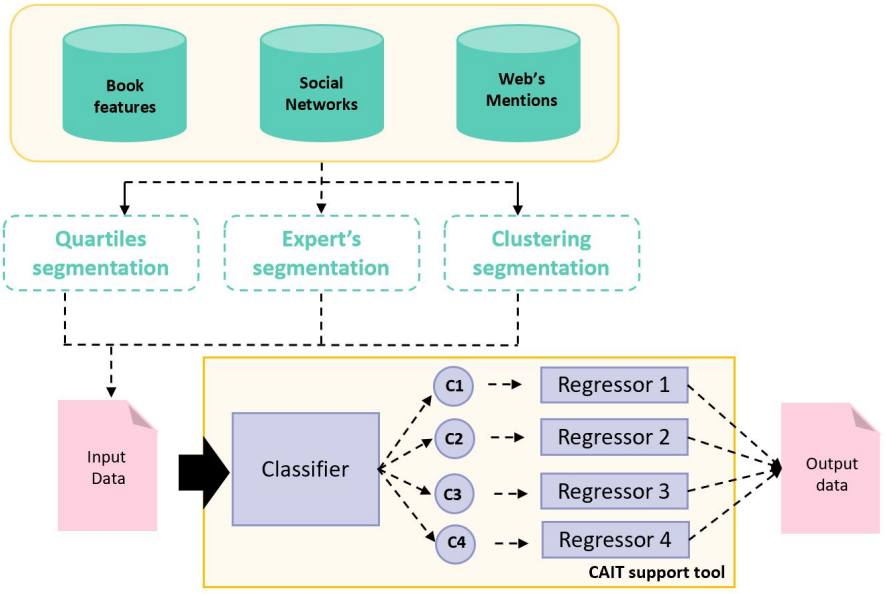

FIGURE 3.7: Research structure diagram. The input data in the combined model of Artificial Intelligence techniques (CAIT) support tool are tested with the three different segmentations. Source: Own elaboration.

3.4 Experimentation algorithms

This section shows the different distributions that the a priori data can have with each of the segmentations to be evaluated. First, the most basic segmentation is performed, quartiles; then expert segmentation is used, and finally, automatic segmentation is given by pattern matching. Subsequently, each of the parts of the combined model is shown. And finally, the evaluation metrics used are exposed.

3.4.1 Segmentations

Each of the segmentations applied to the data before entering the system is described and analyzed below.

Quartiles Segmentation (the Most Basic Segmentation)

The first segmentation to test is the quartiles. For the dataset used, the number of book copies is grouped by an author. In this segmentation, the distribution will be 25% in each of the quartiles, but it cannot be seen correctly in Figure 3.8 a since the existence of outliers is detected, with which they are eliminated above 1.5 of the interquartile range (Q3-Q1). Finally, Figure 3.8b remains, in which the 4 quartiles into which the number of copies can be segmented can be easily identified, being: less than 1808 (Q1—low sales); between 1808 and 4229 (Q2—low intermediate sales); between 4,229 and 12,781 (Q3—high intermediate sales); and finally more than 12781 (Q4—high sales) copies. This last quartile is the so-called Bestseller books. Since the median (Q2) is located in the lower part of the box, it indicates that the data are biased; this is because the quartiles have been selected influenced by the segmentation of the experts after the analysis of the requirements of the Project. We will refer to the quartiles as classes so that later it is easier to compare them with the rest of the segmentations to be analyzed. A better understanding of the distribution of the volume of data by the different segmentations of the quartiles can be seen in Figure 3.9, where an important imbalance of the data is observed between the different segmentations carried out, due to the data biased.

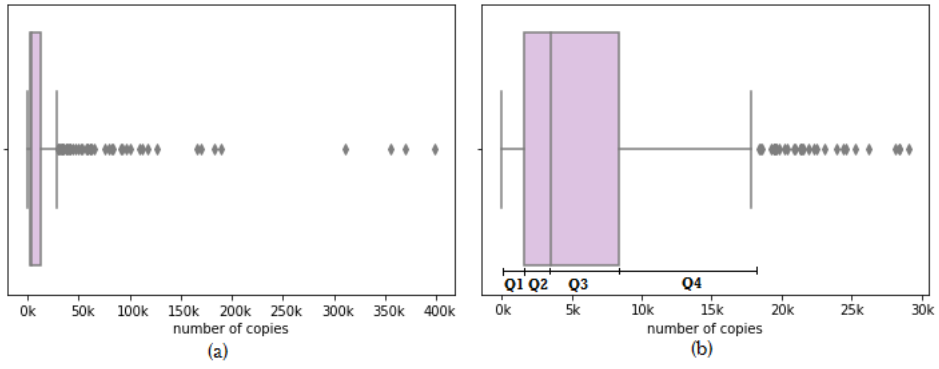

FIGURE 3.8: Results of quartiles segmentation. The boxplot represents the mean (center lines), standard deviation (box), range (dotted lines), and outliers (crosses) of the number of copies of books. (**a**) The quartiles can hardly be appreciated given the number of existing outliers. They are eliminated above 1.5 of the interquartile range, and the quartiles in (**b**) are appreciated where it is observed that in Q1 there will be less than 1808, in Q2 between 1808 and 4229, in Q3 between 4229, and 12 781 and Q4 greater than 12781 numbers of copies. Source: Own elaboration.

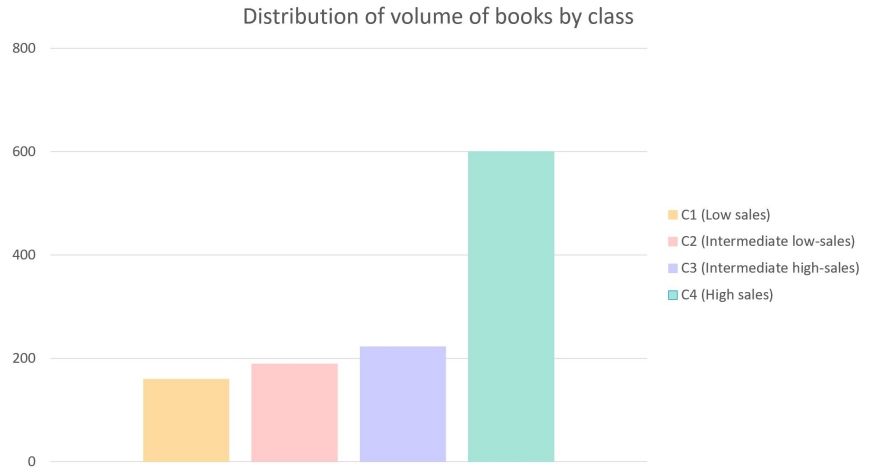

FIGURE 3.9: Distribution of the volume of books for each of the segmentations by quartiles. The results show a significant data imbalance between the different segmentations. Source: Own elaboration.

Expert's Segmentation (the Current Segmentation)

Experts provide the slicers that are currently used by the publisher. They will be identified as class 1 — low sales (C1), class 2 — low intermediate sales (C2), class 3 — high intermediate sales (C3) and class 4 — high sales (C4). The reason for the segmentation carried out by the experts due to confidentiality problems will not be detailed, but Figure 3.10 will show the data volume of the different classes of this segmentation, where the data imbalance continues to exist, especially with respect to the high sales class (C4).

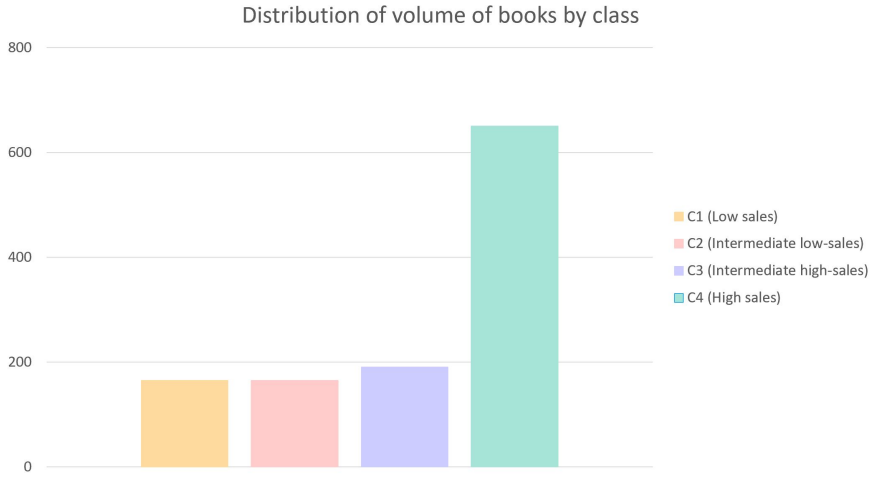

FIGURE 3.10: Distribution of the volume of books for each of the segmentations by experts. The results show a significant data imbalance between the different segmentations. Source: Own elaboration.

Clustering (the Automatic Segmentation) Before applying unsupervised learning techniques for pattern matching, it is necessary to establish the optimal number of k groups. To do this, two different techniques are tested:

- Silhouette index: which consists of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. If most of the objects have a high value, then the cluster configuration is appropriate. If many points have a low or negative value, then the cluster configuration may have too many or too few clusters.
- Elbow curve: which consists of plotting the sum of the squared distances between each point and the centroid in a cluster (Wcss). As the number of clusters increases, the value of Wcss will decrease.

Both evaluation metrics showed as a result a value of $k = 4$ as the number of clusters, as shown in Figure 3.11. These results confirm that the number of segments suggested by the experts from the beginning is correct, but that they can be obtained regardless of their knowledge. To visualize the results of the experimentation with Silhouette index, see Annex A.0.1.

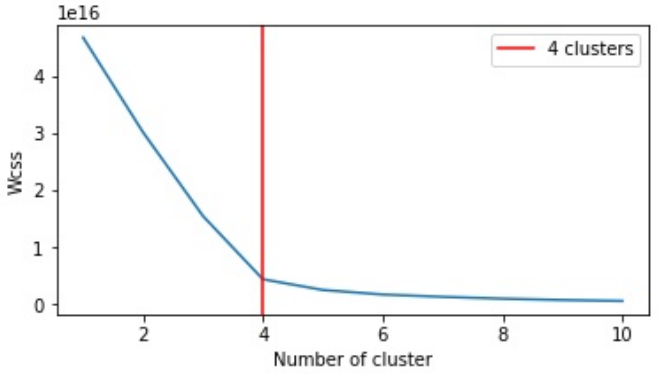

FIGURE 3.11: Elbow curve graph. The blue line indicates the within-cluster sums of squares values. The more this value decreases, the greater the number of clusters. The red line indicates the exact point where the "elbow" occurs, which indicates the optimal number of clusters to choose from, in our case 4. Source: Own elaboration.

Once the optimal number of clusters has been selected to group the data based on their behavior, K-means is used, the simplest and fastest training method. This method only works with numeric data, so the input data include all the variables listed above except (title and author) and genres are converted to numbers, since they represent a single value. Also the dataset has high dimensions, so it will be

impossible to visualize them in order to carry out an analysis. Specifically, there are 49 features of the dataset. If an initial calculation is made, the total number of scatter plots needed to visualize the data would be 49 (features)/2 (number of axes in the plot) = 24. In practice, it is not possible to analyze the data in 24 different plots per for which a dimensionality reduction method is applied with the Principal Component Analysis (PCA) technique for the visualization of the data set [82]. This technique allows us to reduce the number of variables so that we have the minimum number of new variables and that they represent all the old variables in the most representative way possible. Figure 3.12 shows how each of the clusters is grouped.

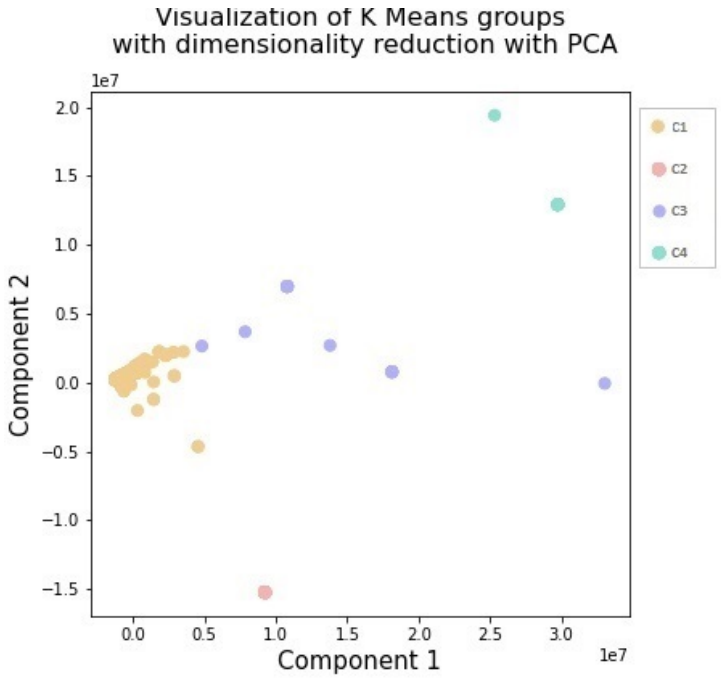

FIGURE 3.12: Visualization of KMeans groups with dimensionality reduction with Principal Component Analysis technique. Source: Own elaboration.

Once the principal components are obtained, the value that each component takes for each observation is calculated based on the original variables, see Figure 3.13. The weights assigned in the first component to the variables corresponding to social networks (lnY, cnT, mentionsNb, etc.) are higher than the sales units and the price, meaning that the first component mainly collects the information corresponding to the interactions in social networks. In the second component, it is the variables (sales units and price) that present the greatest difference in weight, which is why it corresponds mainly to the sales approach. That is why, and being consistent with the previous labels, it is identified as a cluster 1 (C1),

the yellow points; cluster 2 (C2), pink points; cluster 3 (C3), the points of violet color and cluster 4 (C4) the points of turquoise color.

	PC-1	PC-2
units sold	3,100000E-05	2,480000E-03
price	3,100000E-05	2,480000E-03
yearRelease	-4,000000E-06	-2,000000E-06
weekRelease	1,535908E-08	-2,448408E-08
mentionNb	1,770000E-04	-4,520000E-04
cnI	2,242000E-03	2,154000E-03
cnT	2,633659E-08	-9,877190E-09
inT	5,996700E-02	-5,467500E-02
lnT	5,419500E-02	-4,398500E-02
...	...	..
lnY	4,057700E-02	2,263800E-02
fnY	3,355250E-01	1,291250E-01

FIGURE 3.13: Details of each components. Source: Own elaboration.

Figure 3.14 shows the volume of data presented by the 4 clusters detected by KMeans. They will be identified as cluster 1 (C1), cluster 2 (C2), cluster 3 (C3) and cluster 4 (C4), equally an imbalance of the data is observed, but in the natural way that the data is available without having to force it into the segmentations established by the criteria of the experts. In addition, in the figure it can be seen how the criteria of those who belong to the cluster 1 vary drastically, and this is because the criteria that are currently being taken by the experts are arbitrary, and do not really define the nature of the data from the books.

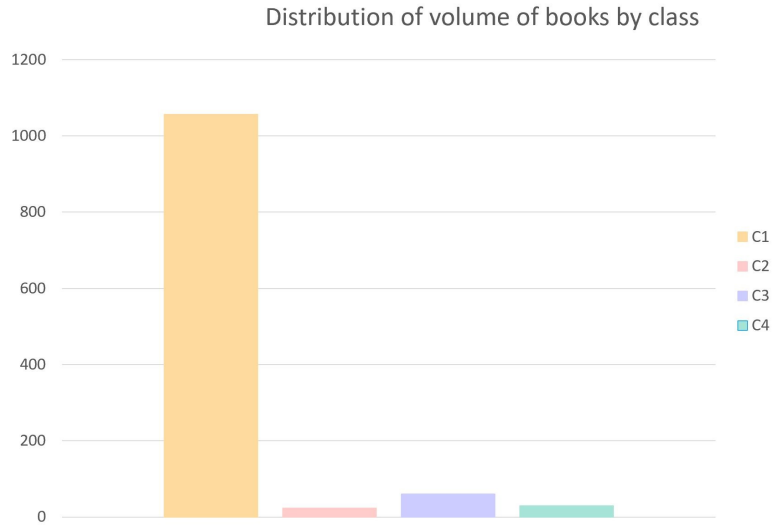

FIGURE 3.14: Visualization of KMeans groups with dimensionality reduction with Principal Component Analysis. Source: Own elaboration.

If the segmentation performed by the experts is compared with the segmentation by clusters in Figure 3.15, a differentiation of data distribution is clearly observed. This is partly due to the fact that the limits established in the current business structure are quite arbitrary, and because it is likely that other factors are influencing that the experts have not been able to share during the development of this research. This point will be covered in depth in Chapter 6.

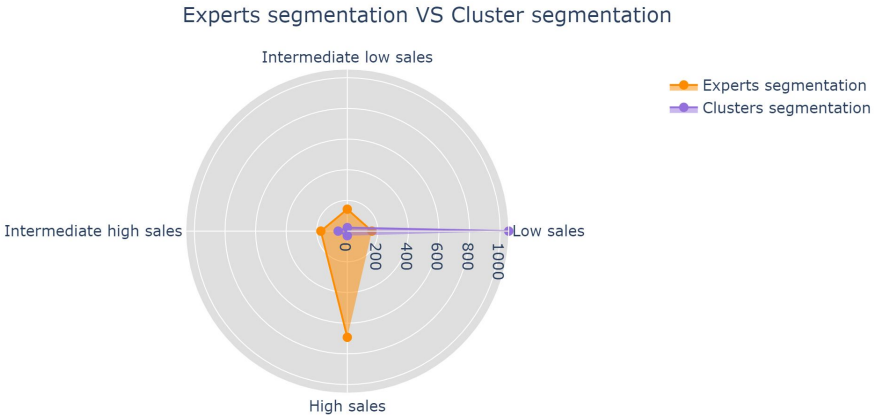

FIGURE 3.15: Comparison of the data distribution for each of the categories between expert segmentations and clusters segmentations. The radius corresponds to the number of cases for each of the categories. Source: Own elaboration.

By having the classes determined by the data patterns themselves, the analysis of each of the classes is deepened. For this, the Shap library [83] (SHapley Additive exPlanations) is used, which is a game theory approach to explain the result of any machine learning model. SHAP is based on these Shapley values, where the objective is to be able to interpret the prediction of a model through the contribution of each input parameter. Likewise, an analysis of each class is performed by the different segmentations: expert segmentation, see Annex A.0.2 and quartiles segmentation, see Annex A.0.3.

For a model where the prediction function is f(x) and F is the set of all input parameters (features) of the model, the Shapley values can be obtained as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{3.2}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the attribution of characteristics for a characteristic j.

In Figure 3.16 the value 8.69 is the prediction of the model for Class 1 according to the characteristics that were input to it. The parameters that are in red are the ones that make the prediction have a higher value, while the parameters that are in blue make the prediction have a lower value. It can be clearly seen that the important variables for purchases turn out to be: `numberPublicationsTW`, `totalReactionsTW`, `numberLikesTW`.

FIGURE 3.16: Force Plot to visualize the importance of each input parameter in the model prediction for Class1. Source: Own elaboration.

In Figure 3.17, belonging to Class 1, it can be seen that the characteristics are ordered by Tree Shape and, therefore, it is observed that despite the fact that the characteristic `value` more globally important, the `totalReactionsTW` feature is the most model-relevant feature in this class; while`numberPublicationsYT` is the least important. The colors red and blue have the same meaning as in the Force Plot. The Figure indicates that the lower the reactions on the social network Twitter, the more potential customers will buy books. Which can be an indication for the marketing department, not to carry out many campaigns in this social network.

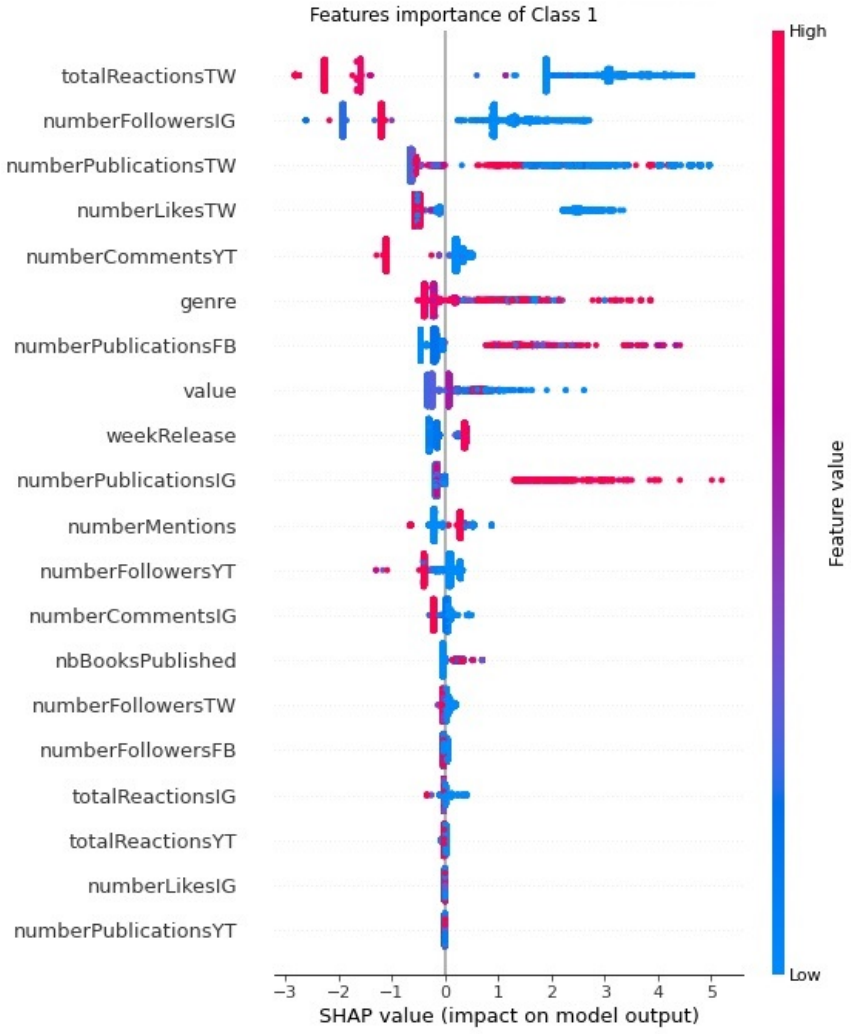

FIGURE 3.17: Summary Plot to see the relationship of all the variables with the model for Class1, and the impact they have. Source: Own elaboration.

If Class 2 is analyzed, it can be seen in Figure 3.18 an effect totally opposite to the previous class. It can be clearly seen that the important variables for purchases turn out to be: `totalReactionIG`, `numberPublicationsIG`.

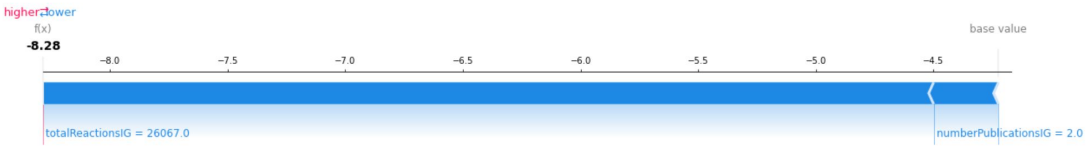

FIGURE 3.18: Force Plot to visualize the importance of each input parameter in the model prediction for Class 2. Source: Own elaboration.

In Figure 3.19, belonging to Class 2, it can be seen that the characteristic `totalReactionsIG` is the most relevant characteristic of the model; while `numberPublicationsTW` is the least important. The colors red and blue have the same meaning as in Force

Plot. From the results shown, it can be intuited that customers will buy more books if there is more activity on the Instagram platform.

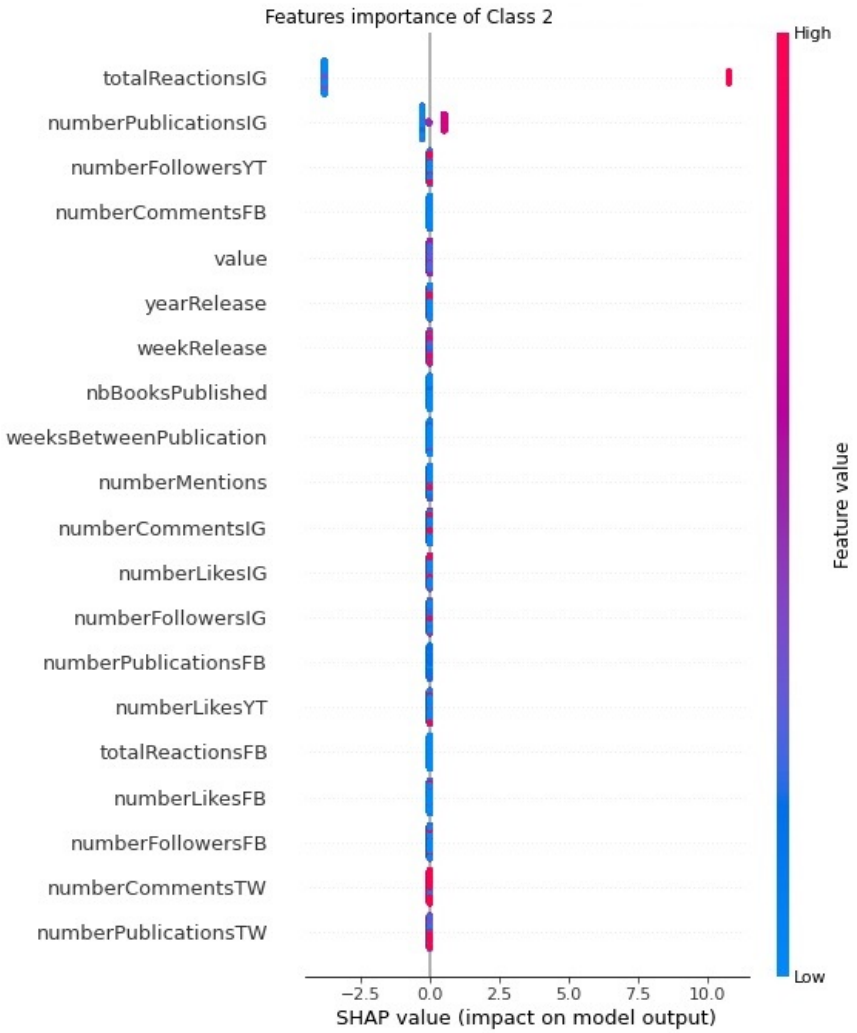

FIGURE 3.19: Summary Plot to see the relationship of all the variables with the model for Class 2, and the impact they have. Source: Own elaboration.

The analysis continues to the next class, where in Figure 3.20 where two characteristics are observed that influence both positively and negatively in the model. The results indicate that the important variables for purchases turn out to be `numberFollowersYT`, `numberFollowersTW`.

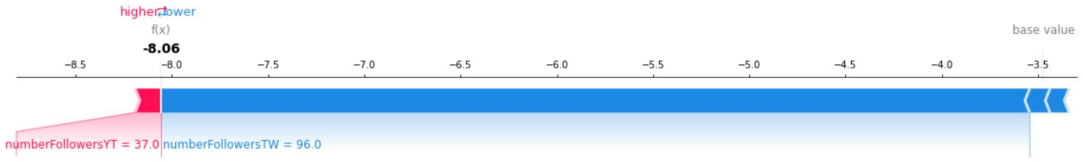

FIGURE 3.20: Force Plot to visualize the importance of each input parameter in the model prediction for Class 3. Source: Own elaboration.

In Figure 3.21, belonging to Class 3, it can be seen that the characteristic `numberFollowersTW`, followed by `numberMentions` are the most relevant characteristics of the model; while `numberLikesFB` and `numberCommentsTW` are the least important. Compared to the books belonging to Class 1, those belonging to this one do show a greater purchase reaction if they interact more on Twitter (especially if they have more followers) and purchases are even influenced by mentions of the authors and their books online.

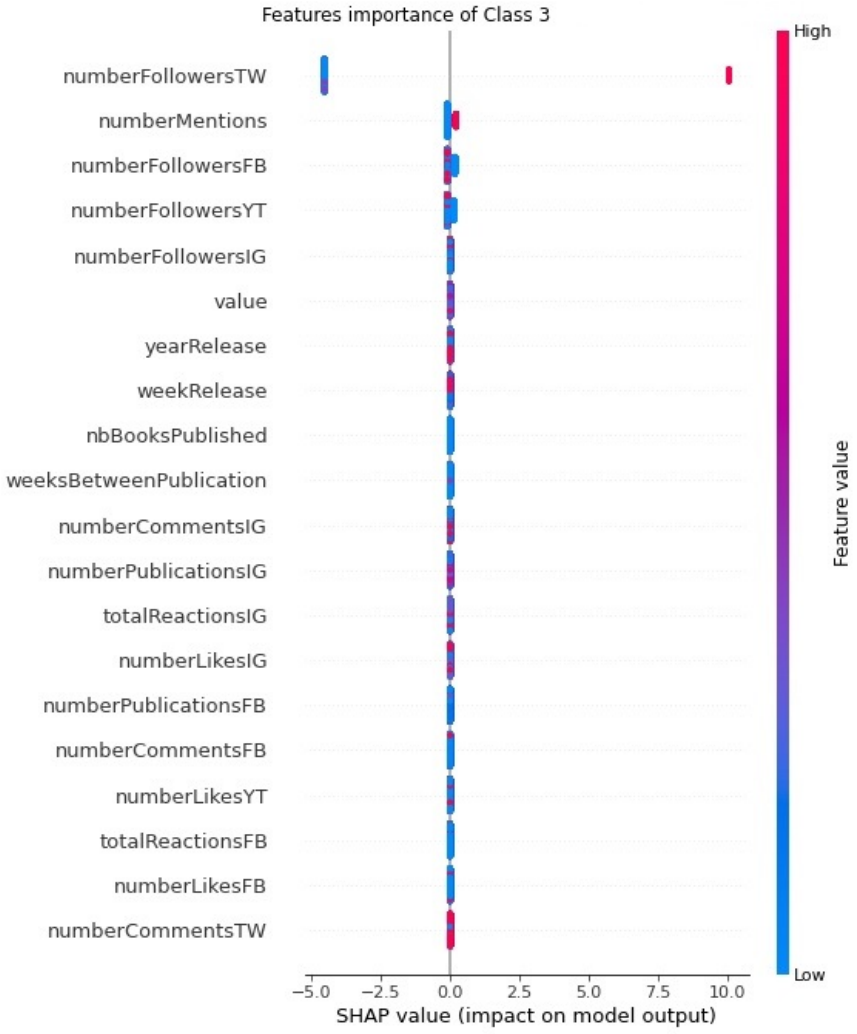

FIGURE 3.21: Summary Plot to see the relationship of all the variables with the model for Class 3, and the impact they have. Source: Own elaboration.

Finally, Figure 3.22 shows that the important variables for purchases turn out to be `numberFollowersYT` and `weekRelease`. Also, show that the variable `weekRelease` has a positive effect, which makes the prediction have a higher value.

FIGURE 3.22: Force Plot to visualize the importance of each input parameter in the model prediction for Class 4. Source: Own elaboration.

In Figure 3.23, belonging to Class 4, it can be seen that the characteristic `numberFollowersYT`, followed by `weekRelease` are the most relevant characteristics of the model; while `numberCommentsTW` and `numberPublicationsTW` are the least important. However, in the last class it can be seen that the platform that most influences purchases is YouTube (especially the number of followers). Another variable that influences purchases is the week the book is published. The more weeks there are between book launches, the more purchases will be made.

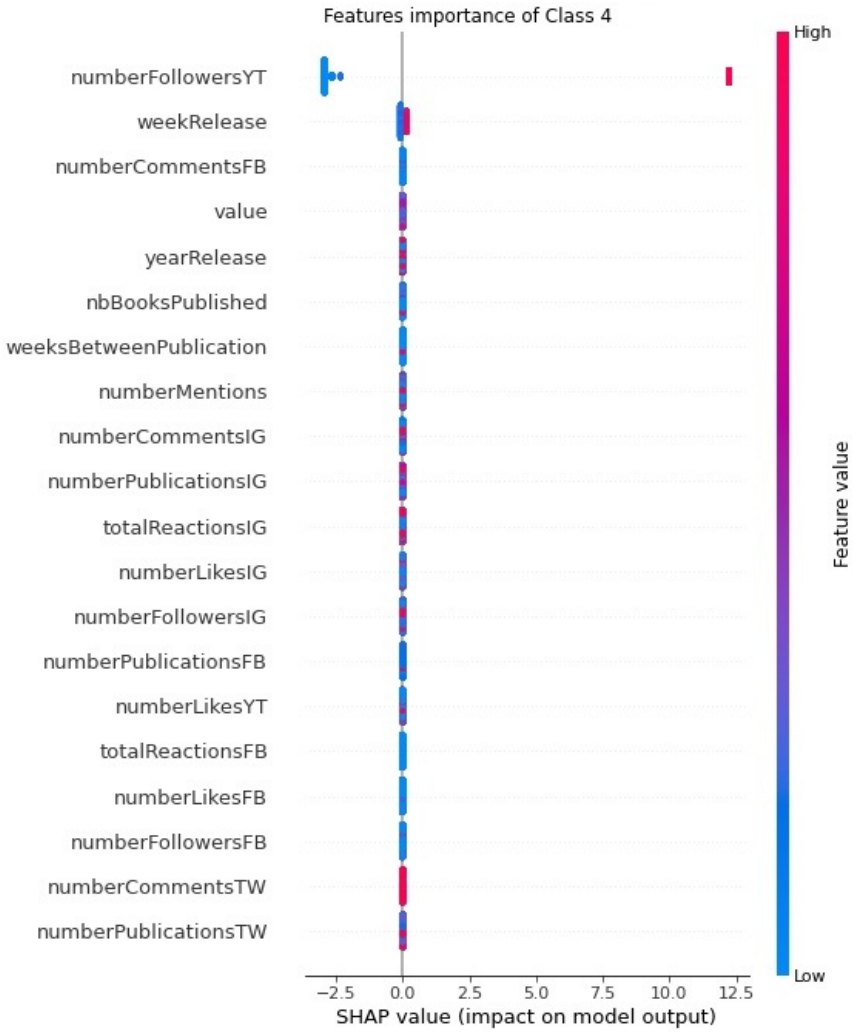

FIGURE 3.23: Summary Plot to see the relationship of all the variables with the model for Class 4, and the impact they have. Source: Own elaboration.

3.4.2 Part A: The Classifier

Four classification algorithms shall be considered for the implementation of the classifier. These are:

- Decision Tree: It is a representation in the form of a tree whose branches branch according to the values taken by the variables and which end in a specific action. It is generally used when the number of conditions is not very large in this study. See [84–86] for a detailed description of this algorithm.
- Random Forest: It is a combination of a predictor trees such that each tree depends on the values of a random vector tested independently and with the same distribution for each of these. It is implemented in data mining to classify or forecast a target variable. See [87–89] for a detailed description of this algorithm.
- K-Nearest Neighbors: It is a classification method used to estimate the density function of the predictors for each class. See [90–92] for a detailed description of this algorithm.
- XGBoost: Part of the decision tree that is implemented in data mining to classify or forecast on a target variable (book copies), through machine learning that is performed on a set of data, using several weak classifiers. In this case, they are the decision trees, but enhancing the results of these, due to the sequential processing of the data with a loss or cost function, minimizes the error iteration after iteration, thus making it a strong predictor. However, this will depend on the level of adjustment of the parameters used in the function. See [93–95] for a detailed description of this algorithm.

For the specific dataset working, as will be detailed in the next section, the best algorithm is XGBoost. The implementation of this algorithm can be seen in Algorithm 1, in which the input data are entered that is made up of the book's characteristics, the author's social network data, plus the mentions it has on the web (X_i). Finally, the output is given by the different segmentations into which books can be divided based on the number of copies of a book (Y_i).

Algorithm 1 Classifier phase

Split: D_{total} in D_{train} and D_{test} (from K-Fold stratified cross-validation, in this case k = 10)

Input: $D_{train} = (X_j, Y_j)$ Where the target variable will be the segmentation

1: An initial tree F_0 is obtained to predict the objective variable Y_j, the residual is associated with the difference $(Y_j - F_0)$.

2: A new tree "h_1" is obtained that adjusts the error to the previous weight.

3: The results of F_0 and h_1 are combined to obtain the tree F_1, where the mean square error of F_1 will be less than that of F_0

$$F_1 x < -F_0 x + h_1(x)$$

4: This process is continued iteratively until the error is minimized as much as possible in the following way:

$$F_m x < -F_m - 1x + h_m(x)$$

5: The classifier is tested with D_{test} using Accuracy, Precision, Recall, F1Score, and MAE as the evaluation metrics.

Output: Y_{class}, predicted segmentation.

3.4.3 Part B: The Regressors

In the second part, we use the regressor algorithms; with them, the aim is to study the effect of one or more independent variables on a single dependent variable. The dependent variable (Y) will be the one we seek to survey through statistical regression to understand how it adapts when modifying the independent variables (X_i). After mathematically describing what has just been explained, it can be obtained in Equation 3.3.

$$Y = 0 + B_1 * X_1 + B_2 * X_2 + \ldots + B_n * X_n + \varepsilon \tag{3.3}$$

where Y represents the dependent variable that is being studied or trying to predict, $X_1, X_2 ... X_n$ are all the independent variables that influence or can affect the dependent variable Y. The function of ϵ is to explain the possible variability of the data that cannot be presented through the linear relationship of the formula; in other words, it represents the possible existing error.

Knowing the objective and operation of the regressive algorithms, the ones selected for the competition are shown below, looking for the one that best suits the input data:

- Gradient Boosting: It is a machine learning technique [96–98] which produces a predictive model in the form of a set of weak prediction models (typically decision trees). When building the model, it is done in a stepwise manner (as boosting methods do), and it generalizes them, allowing the arbitrary optimization of a differentiable loss function.

- XGBoost: Described in a previous section.
- LightGBM: It is a distributed gradient impulse framework for machine learning. It is based on decision tree algorithms but does not grow at the tree level but in leaves. Therefore, by choosing the one will produce the greatest decrease in loss. See [99, 100] for a detailed description of this algorithm.

For the specific dataset working, as will be detailed in the next section, the best algorithm is XGBoost. The implementation of this algorithm can be seen in Algorithm 2, in which the characteristics of the book, the data of the author's social network, plus the mentions it has on the web (X_j). The output is the objective variable that will be given by the number of copies (Y_j).

Algorithm 2 Regressor phase

Split: D_{total} in D_{train} and D_{test} (from K-Fold stratified cross-validation, in this case $k = 10$)

Input: $D_{train} = (X_j, Y_j)$ Where the target variable will be the number of copies

1: An initial tree F_0 is obtained to predict the objective variable Y_j, the residual is associated with the difference $(Y_j - F_0)$.
2: A new tree "h_1" is obtained that adjusts the error to the previous weight.
3: The results of F_0 and h_1 are combined to obtain the tree F_1, where the mean square error of F_1 will be less than that of F_0

$$F_1 x < -F_0 x + h_1(x)$$

4: This process is continued iteratively until the error is minimized as much as possible in the following way:

$$F_m x < -F_m - 1x + h_m(x)$$

5: The regressor is tested with D_{test} using R^2 as the evaluation metric.

Output: $Y_{predicted}$, predicted number of copies.

3.4.4 CAIT: A Predictive Support Tool

Once each component of the proposed predictive support tool has been described, its implementation can be observed in Algorithm 3, where the characteristics of the book, the author's social network, web mentions and the segmentation of the book are introduced in said data (X_k). This data will go through the classification function, obtaining as a result which segmentation group each book can belong to. Given this classification, the data corresponding to each trained group will be divided, and the regressors will be applied individually through hyperparameter optimization. The detail of the hyperparameters used can be seen in the Results section.

Algorithm 3 CAIT algorithm

Input:X_k

```
class = Classifier phase (X_k)
if class == 1:
    Regressor phase (X_k)
elseif class == 2:
    Regressor phase (X_k)
elseif class == 3:
    Regressor phase (X_k)
elseif class == 4:
    Regressor phase (X_k)
```

Output: $Y_{nbcopies}$, number of book copies to print according to its segmentation in the market.

Finally, in this subsection, the composition of the predictive support tool created has been detailed. Demonstrating the effect of getting the best of each part and joining them into one helps improve copy number accuracy.

3.4.5 Evaluation metrics

The evaluation metrics will allow to measure the quality of the machine learning model in classification tasks. To do this, many of the metrics described below use positive values (TP), which describe the rate at which the classifier predicts observations that are "positive" as "positive." False Positive (FP) values describes the rate at which the classifier predicts observations that are actually "negative' as "positive". Negative values (NT), which describe the rate at which the classifier predicts observations as "negative". observations that are "negative". False negative (FN) values describe the rate at which the classifier predicts observations that are actually "positive" as "negative".

For a better understanding of the reader of this thesis, a confusion matrix is presented which verifies how well the classifier algorithm performs and explains in a visual way the values (TP,FP,TN,FN) that will be used in the equations. of the evaluation metrics. If you observe Figure 3.24 for a binary problem, for example if I like a book or not, it can be seen that 226 people really do like the book, and yet 30 have been classified as not they like it, when really it is the opposite.

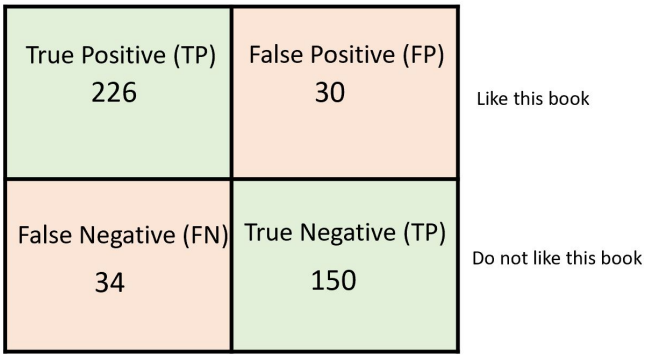

FIGURE 3.24: Confusion matrix for a binary problem. Source: Own elaboration.

For this investigation, the algorithm will order the books according to the number of copies sold in 4 classes (C1,C2,C3,C4). This type of classification is called multiclass classification. Once understood, these concepts are described below the evaluation metrics used for this research:

- K-Fold stratified cross-validation, this validation seeks to ensure that each k group is representative in all data strata. It is intended to ensure that each class is (roughly represented equally in each test fold) and thus avoid overtraining. In this specific case, the variable $k = 10$.
- *Accuracy* (Equation (3.4)), which refers to how close a sample statistic is to a population parameter.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

  where TP, represent true positive values, TN, represent true negative values, FP, represent false positive values, FN, represent false negative values.

- *Precision* (Equation (3.5)), quantifies the number of positive class predictions that actually belong to the positive class.

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

- *Recall* (Equation (3.6)), represents the model's ability to correctly predict the positives out of actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

- *F1-Score* (Equation (3.7)), this gives a weighted average of the precision and recall metrics. It is the best metric for averaging out and balancing all the evaluation metrics as a whole.

$$F1\text{-}Score = \frac{Precision * Recall}{Precision + Recall} \tag{3.7}$$

- R^2 (Equation (3.8)) is a statistical metric in the regression models that allows determining the proportion of variance in the dependent variable, which is explained by the independent variable.

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} \tag{3.8}$$

where $SS_{regression}$ is the sum of squares due to the regression *(sum of squares explained)* and SS_{total} is the total sum of squares.

3.4.6 Analysis of the sales cycle by groups

Once the sales groups have been determined, the sales model they follow is analyzed, with the aim of developing actions and strategies to increase them. For this, five standardized models are applied in the field of marketing:

Bass model: It consists of a simple differential equation that describes the process of how new products are adopted in a population. The model presents a rationale for how current adopters and potential adopters of a new product interact [101]. The formula that these model uses is Equation 3.9.

$$F_i(t) = pv^{\infty} \left[\frac{1 - e^{-(p+q)^t}}{1 + \frac{q}{p} e^{-(p+q)^t}} \right] \tag{3.9}$$

where p characterizes "innovators", reflecting an influence that is independent of actual book units sold (pv_i^t), and q reflects the imitation part of the model.

Barabasi Model: This statistical model [43] studies the evolution of the sales of best-selling books. It is mainly based on the following three basic mathematical concepts.

- The physical factor of the book: The suitability of a book η_i is a concept that approximates the ability of the book to respond to the taste of a wide audience. This ability is influenced by the writing style, the illustrations inside and on the cover, the genre, the time, and many other factors.

- The preferential attachment factor of the book: It is a concept that refers to the greater probability that a book will be bought / read when this book already has a very good sales record and a very good rating among critics and the public. Preferential attachment is a "social" factor that, translated into mathematics, means that the probability of buying a book depends on its current sales, S_i^t. These sales of a book at a given time are defined in the Equation 3.4.6:

$$S_i^t = m\left[e^{\lambda_i \Phi(\frac{\ln t - \mu_i}{\sigma_i})} - 1\right],$$

  where:

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2}\, dy$$

  is the cumulative normal distribution; m is the average number of references or citations that a new article contains. The authors found that this parameter does not affect the results and they decided to set it to m = 30; λ_i is the relative adequacy that captures the importance of an item compared to other items; Φ is the previously mentioned immediacy factor of an article and μ_i is the longevity factor of a previously mentioned article, capturing its citation decay rate.

- The aging factor of the book: It is a concept that models the gradual loss of interest that the public has in that book, after its release date. It is defined as the Equation 3.4.6:

$$A_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} \exp\left[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}\right]$$

  In this equation there are two important parameters that, together with the adequacy factor η_i, will be calculated and adjusted to obtain the book sales function. One is the immediacy factor μ_i, which represents the time when sales peak. The other is the decay rate σ_i, which represents the longevity of the book.

From these three points, and combining their equations, it can be said that the probability $\Pi_i(t)$ of the book i that will be purchased at time t after the publication date is:

$$\Pi_i(t) \sim \eta_i S_i^t A_i(t)$$

Relative suitability alone can predict how many copies a book will sell during its lifetime. Taking $t \to \infty$ into the sales equation gives:

$$S_i^{\infty} = m\big(e^{\lambda_i} - 1\big)$$

predicting that the total number of sales of a book in its lifetime depends only on a single parameter, the relative fitness λ.

Logistics Model: The logistic function is widely used to model population growth and product adoption, with applications in many fields. In the context of books, one might view a recently published book as a new product, the adoption of which leads to an increase in sales. [102] The formula used by this model corresponds to Equation 3.10.

$$F_i(t) = \frac{uv_i^{\infty}}{1 + e^{-r_i(t-\tau_i)}}, \tag{3.10}$$

where uv_i^{∞}, r_i and τ_i correspond to the final sales, longevity and immediacy of a book i.

Exponential Model: This statistical model [103] is based on three main parameters:

- Maximum Book Units Sold: This A parameter is related to the highest maximum book units sold.
- The growth factor: This parameter α represents the growth factor of the book units sold over time. Govern the first ascending part of the sales-time-units function.
- The decomposition factor: This parameter, β, represents the gradual decrease in unit sales over time. The second part governs, which falls from the sales-time units function.

The general function that defines the total accumulated book units sold at a given time is Equation 3.11.

$$F_i(t) = \frac{A}{\alpha_i - \beta_i}\left[\big(\frac{1}{\alpha_i}\big)(e^{-\alpha_i t} - 1) - \big(\frac{1}{\beta_i}\big)(e^{-\beta_i t} - 1)\right] \tag{3.11}$$

If t $\to \infty$, the total final units that the book would get, according to this exponential model, is defined by Equation 3.12.

$$f_i(t) = \frac{A}{\alpha_i - \beta_i}\big[(e^{-\beta_i t}) - (e^{-\alpha_i t})\big] \tag{3.12}$$

Gompertz Model: The Gompertz model [104], named for Benjamin Gompertz, was first proposed to model human mortality. The model generates a skewed diffusion curve with long tails. In this context, the first books sold will pave the way for the sales of the next ones and boost your sales dynamics. The formula that these model uses is Equation 3.13.

$$F_i(t) = uv_i^{\infty} * e^{-e^{-(a_i + q_i * t)}}, \tag{3.13}$$

where uv_i^{∞}, a_i and q_i correspond to the books sold, the displacement in uv_i^t and the sales growth rate, respectively.

3.5 Results

All the values of the metrics discussed above are combined into a single overall scorecard for each classifier for each of the segmentations. Before applying classification techniques, as previously observed in Figures 3.11, 3.12 and 3.14, the data is unbalanced. Which modifies the original distribution of the sample either by eliminating cases or instances of the majority class (Undersampling) or by replicating or creating new instances of the minority class (Oversampling). In this experimentation, it is specifically used in Oversampling (specifically RandomOverSampler). This data balancing technique is applied to those classes that consist of a reduced number of samples (without eliminating real data, as other techniques do), but which constitute the main objective of the classification process.

The XGBoost classifier algorithm outperforms the decision tree, KNN and random forest in all results, as can be seen from Tables 3.2 and 3.3. Observing Table 3.4, KNN could also be selected, but XGBoost is chosen as the optimal one, given its flexibility as a parallelize algorithm, and that in general in the three segmentations, the scores of this algorithm are higher than other classifiers. The selected classifier can provide very good consistency across classes. Analyzing the results obtained from the segmentation of the expert (Table 3.3) with respect to the segmentation of quartiles (Table 3.2) there is no great difference with what can be interpreted, since that the use of this type of segmentation does not improve or worsen the expert's segmentation. However, if the expert segmentation is compared with the clustering segmentation (Table 3.4), it is observed that the classification improves significantly since it fully adjusts to the patterns found in the data.

TABLE 3.2: Comparison of classifiers algorithms with quartiles segmentation. This table shows that the XGBoost algorithm is the optimal one of the four to be compared since it presents the best accuracy, with the lowest mean absolute error.

	Decision Tree	**K-Nearest Neighbors**	**Random Forest**	**XGBoost**
Accuracy	0.86	0.86	0.89	**0.91**
Precision	0.86	0.86	0.89	**0.91**
Recall	0.86	0.86	0.89	**0.91**
F1-Score	0.86	0.86	0.89	**0.91**

TABLE 3.3: Comparison of classifiers algorithms with expert's segmentation. This table shows that the XGBoost algorithm is the optimal one of the four to be compared, since it presents the best accuracy, with the lowest mean absolute error.

	Decision Tree	**K-Nearest Neighbors**	**Random Forest**	**XGBoost**
Accuracy	0.89	0.87	0.90	**0.93**
Precision	0.89	0.87	0.90	**0.93**
Recall	0.89	0.87	0.90	**0.93**
F1-Score	0.89	0.87	0.90	**0.93**

TABLE 3.4: Comparison of classifiers algorithms with clustering segmentation. This table shows that both the XGBoost and KNN algorithms can be the most optimal of the four to be compared, given that they present the best accuracy, with the lowest mean absolute error.

	Decision Tree	**K-Nearest Neighbors**	**Random Forest**	**XGBoost**
Accuracy	0.99	**1.00**	0.99	**1.00**
Precision	0.99	**1.00**	0.99	**1.00**
Recall	0.99	**1.00**	0.99	**1.00**
F1-Score	0.99	**1.00**	0.99	**1.00**

Once they are classified into the 4 groups: class 1 (C1) corresponds to low sales and/or cluster 1, class 2 (C2) to low intermediate sales and/or cluster 2, class 3 (C3) to high medium sales and/or cluster 3, and finally class 4 (C4) to increased sales and/or cluster 4. The prediction of the specimens for each of the classes is determined.

Hyperparameters of the classifiers used

To find the most optimal hyperparameters for each classification algorithm, GridSearch is used, a search method that takes into account different combinations of hyperparameters and chooses the combination that yields the lowest margin of error.

- Decision Tree: max leaf nodes=50, min samples split= 2, random state = 0
- Random Forest: n estimators = 100, criterion = gini, min samples split= 2, random state = 0
- K Nearest Neighbors: n neighbors=20
- XGBoost: booster= gbtree, eta= 0.3, max depth= 4, subsample=0.6, min child weight= 2

The optimal algorithm is the XGBoost Regressor, as can be seen again, since it is the one with the best results of the three. If a comparison is made between the different segmentations, it can be seen in Table 3.5 that the segmentation by quartiles presents the lowest prediction values. In Table 3.7, it is clearly observed that the prediction with segmentation by clusters is very similar to the results with segmentation by experts, as can be seen in Table 3.6. The real difference between the two results is that the data from the experts will be previously labeled with the classification they establish, however with clustering segmentation, the machine itself will find the corresponding label through the patterns. This is of vital importance, since when a new book arrives, the machine itself will be able to label which class it corresponds to.

TABLE 3.5: Comparison of regressors algorithms using R^2 as the evaluation metric. This table shows that the XGBoost algorithm is the one that best predicts the number of copies of books for each of the segmentations obtained with segmentation by quartiles.

	Class 1	**Class 2**	**Class 3**	**Class 4**
GBoosting	0.75	0.72	0.51	0.16
XGBoost	**0.93**	**0.96**	**0.98**	**0.94**
LGBM	0.87	0.72	0.51	0.43

TABLE 3.6: Comparison of regressors algorithms using R^2 as the evaluation metric. This table shows that the XGBoost algorithm is the one that best predicts the number of copies of books for each of the segmentations obtained with expert's segmentation.

	Class 1	**Class 2**	**Class 3**	**Class 4**
GBoosting	0.32	0.34	0.77	0.06
XGBoost	**0.94**	**0.96**	**1.00**	**0.96**
LGBM	0.20	0.00	0.00	0.38

TABLE 3.7: Comparison of regressors algorithms using R^2 as the evaluation metric. This table shows that the XGBoost algorithm is the one that best predicts the number of copies of books for each of the segmentations obtained with the clustering segmentation.

	Class 1	Class 2	Class 3	Class 4
GBoosting	0.73	0.73	0.13	0.16
XGBoost	**0.95**	**0.97**	**1.00**	**0.96**
LGBM	0.90	0.87	0.86	0.41

Hyperparameters of the regressors used

For the GBoosting and LGBM regression algorithms in each class the default parameters are used. In the case of the algorithm selected as optimal (XGBoost), the hyperparameters optimized for each class are described below:

- Regressor 1: lambda = 3; booster = gblinear, alpha = 5, feature selector = shuffle
- Regressor 2: lambda = 5; booster = gblinear, alpha = 18, feature selector = cyclic
- Regressor 3: lambda = 4; booster = gblinear, alpha = 12, feature selector = cyclic
- Regressor 4: lambda = 8; booster = gblinear, alpha = 2, feature selector = shuffle

For a better understanding of the reader of this thesis, Annex A.0.4 offers a sample of the predictive results of the number of copies of books (without and with the established classes). And Annex A.0.5 offers a visualization of the report of the number of books that will be sold by each new author.

Below 3.25 is displayed with a sample of 50 data, the actual values of units and the predicted value. Although there are points that may not be adjusted, the vast majority have a fairly significant success.

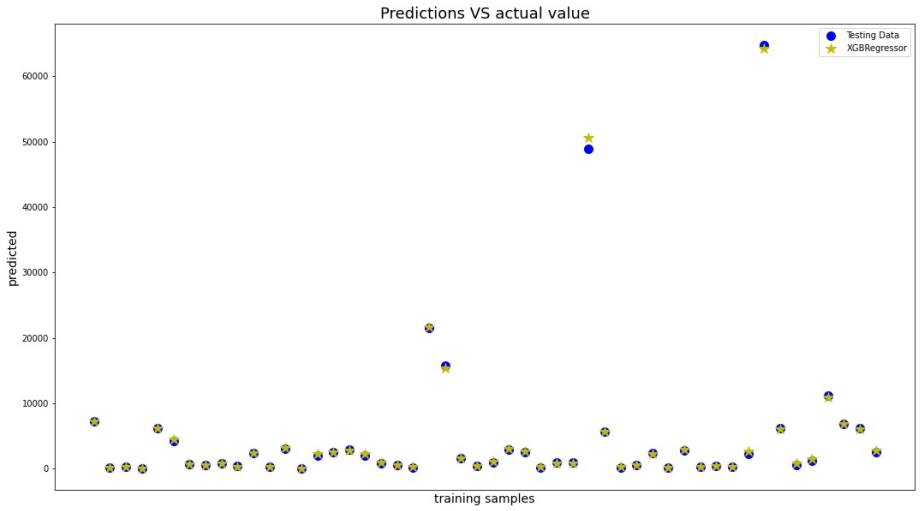

FIGURE 3.25: Predictions versus actual value of sample of 50 data. Source: Own elaboration.

Analysis of the sales cycle by groups

In order to analysis of the sales cycle of each group segmented by clusters (the one that provides the best results according to the above), the five statistical models mentioned in the Algorithm Experimentation section are initially tested with all the data. And once the sales cycle to follow has been selected, it is analyzed by each group.

For the sales analysis, the estimation of the residual sum of squares (SSE) is used as an evaluation metric. This last metric corresponds to a measure of discrepancy between the data and the estimation models, indicating that if a lower value is obtained, the models fit perfectly. Their results can be seen in Table 3.8. The determination coefficient (R^2) is also evaluated, which can acquire a value between 0 and 1, indicating that 1 would be the best fitted model, its results are observed in Table 3.9.

TABLE 3.8: Comparison of statistical models using SSE as the evaluation metric. This table shows that the best results are presented by both the Exponential model and the Barabasi model.

Model	**mean SSE**	**stdDev SSE**
Bass	3.354879e+07	2.220836e+08
Barabasi	**3.166727e+07**	**2.148118e+08**
Gompertz	3.412491e+07	2.202430e+08
Logistic	5.113655e+07	3.009369e+08
Exponential	**2.892330e+07**	**1.980244e+08**

TABLE 3.9: Comparison of statistical models using R^2 as the evaluation metric. This table shows that the best results are presented by both the Bass, Barabasi and Exponential models.

Model	**mean R^2**	**stdDev R^2**
Bass	**0.037473**	**0.160959**
Barabasi	**0.012270**	**0.700270**
Gompertz	-11.243984	196.854925
Logistic	-16.010281	213.303608
Exponential	**0.047643**	**0.174125**

The results show that the determination coefficient metric provides a fairly low value, so another evaluation metric is applied to have enough resources for the conclusions, in this case the chi^2 test is used. With this test from the significance factor called alpha (in this case with a value of 0.05) the model or models to be used are determined. If the result obtained is greater than the alpha value, the hypothesis will be null, in this case, the model will be discarded and vice versa. The results can be seen in Table 3.10.

TABLE 3.10: Comparison of statistical models using chi^2 as the evaluation metric. This table shows that the best results are presented by both the Barabasi and Exponential models.

Model	**mean chi^2**	**stdDev chi^2**
Bass	0.052553	0.200214
Barabasi	**0.041386**	**0.180579**
Gompertz	0.050981	0.1962915
Logistic	0.056543	0.211354
Exponential	**0.025189**	**0.124447**

Globally, analyzing the three evaluation metrics, it can be seen that they coincide in the same type of model that will allow us to analyze the sales cycle, which are: the Barabasi model and the Exponential model. Bearing in mind that the best segmentation (and the one that will be selected for the sales analysis) was segmentation by clustering; Figures 3.26, 3.27, 3.28 and 3.29 show the analysis for each group, distinguishing the sales cycles that identify it. In most segments, they are governed by the Barabasi sales cycle, which indicates that a book is most likely to sell when it has achieved record sales and a good rating from book critics and the public. Exceptionally, in Figure 3.29 it can be seen that a subject (youth literature) follows an Exponential sales cycle; however in Figure 3.26 where there is a higher volume of topics (exactly 16), they follow the exponential sales model.

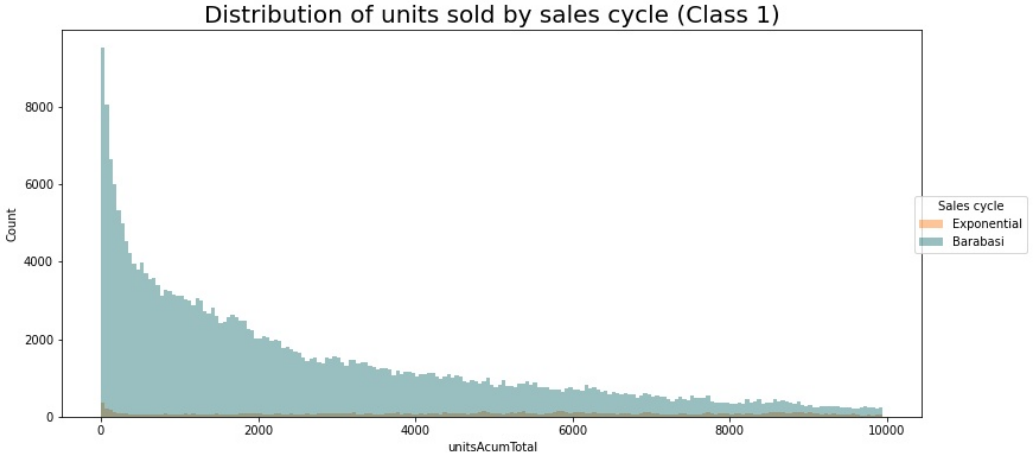

FIGURE 3.26: Distribution of units sold by sales cycle (Class 1). It shows us that the great majority of the books of this class will respond to the sales of the Barabasi model, which refers to the greater probability that a book will be bought when this book already has a certain sales record and a very good rating between critics and the public. Of these, 16 themes respond to the Exponential sales cycle. Source: Own elaboration.

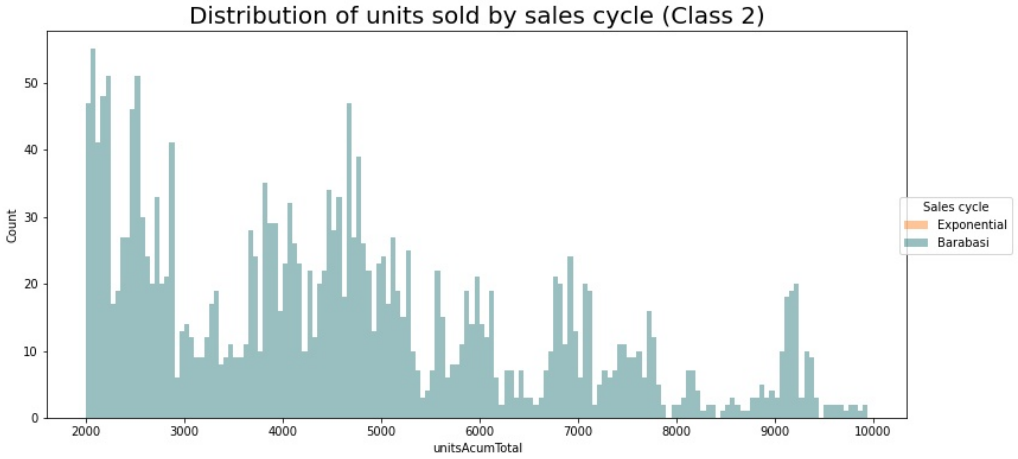

FIGURE 3.27: Distribution of units sold by sales cycle (Class 2). It shows us that the great majority of the books of this class will respond to the sales of the Barabasi model, which refers to the greater probability that a book will be bought when this book already has a certain sales record and a very good rating between critics and the public. Source: Own elaboration.

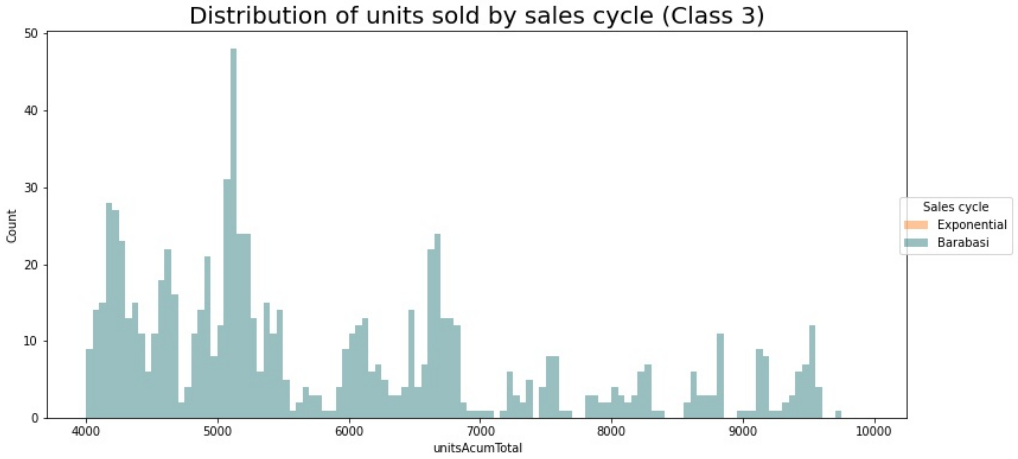

FIGURE 3.28: Distribution of units sold by sales cycle (Class 3). It shows us that the great majority of the books of this class will respond to the sales of the Barabasi model, which refers to the greater probability that a book will be bought when this book already has a certain sales record and a very good rating between critics and the public. Source: Own elaboration.

In these figures also allow observing the range of values of each of the sales groups (accumulated units) obtained from the grouping segmentation.

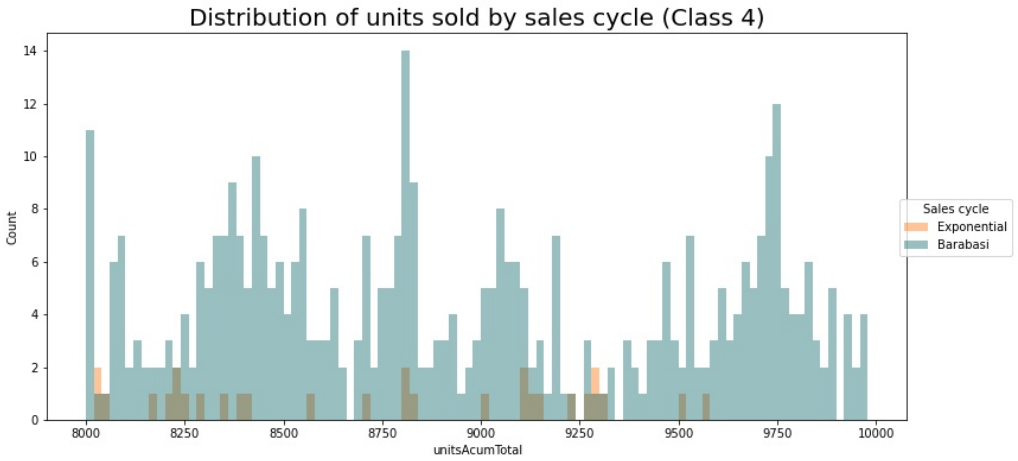

FIGURE 3.29: Distribution of units sold by sales cycle (Class 4). It shows us that the great majority of the books of this class will respond to the sales of the Barabasi model, which refers to the greater probability that a book will be bought when this book already has a certain sales record and a very good rating between critics and the public. Only one theme respond to the Exponential model: 'LITERATURA JUVENIL' (YOUTH LITERATURE in English). Source: Own elaboration.

Since the first class has a larger volume of books, Figure 3.30 shows all the existing topics and the sales cycle to which each one corresponds.

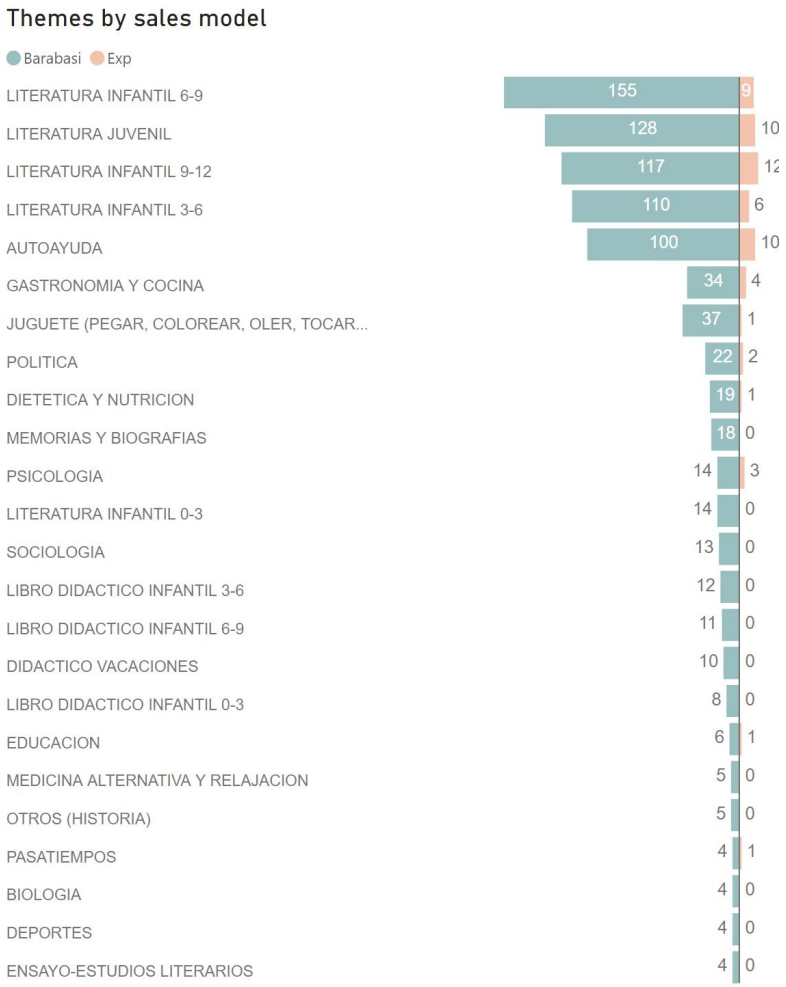

FIGURE 3.30: Themes by sales cycle. Source: Own elaboration.

With the image above, it can be seen that the vast majority of issues in the Non-fiction and Children's/Youth genres manage to reach a rapid sales record in a short period, thus responding to the Barabasi model, and rarely present a cycle of sales in which they achieve their maximum value over time. If the first 5 topics are analyzed, it can be seen that the ones that sell the fastest and achieve high sales are those of children's literature in the range of 3 to 12 years and youth, this shows some agreement with real life, since it is in this age group where both parents and schools reinforce reading the most. It is also noted that self-help books are another subject that sells out quite quickly, influenced by criticism. Both in this topic and in the previous ones, it is observed that there are some of them whose sale manages to reach maximum sales over time, with the Exponential cycle that they present.

Figure 3.31 shows an example of a book that fits one of the models, in this case, the book ean/isbn 9781474928601 (blue line) fits the Barabasi model (red dashed line) and Figure 3.32 shows an example of a book that fits one of the models, in this case, the book ean/isbn 9788427044784 (blue line) fits the Exponential model (red dashed line).

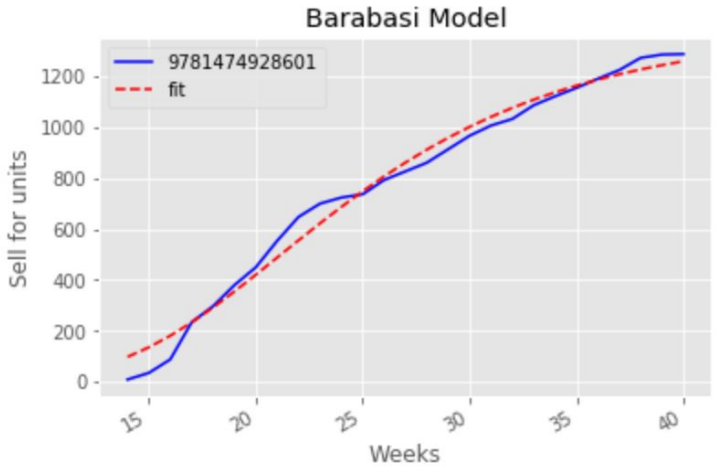

FIGURE 3.31: Adjustment of book sales to the Barabasi model. Source: Own elaboration.

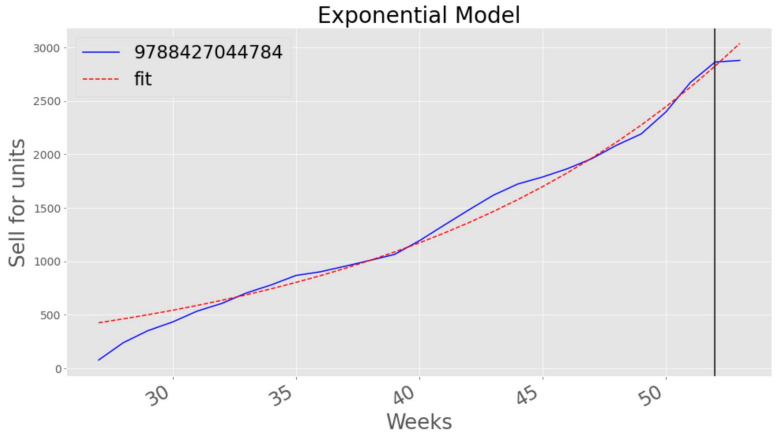

FIGURE 3.32: Adjustment of book sales to the Exponential model. Source: Own elaboration.

3.6 Summary

In this section, different Artificial Intelligence techniques have been experimented with, until finding the optimal solution to adjust the predictions of copies of Non-fiction and Children's/Youth books. It has been shown that their combined use, with segmentation based on behavior patterns, provides better results in terms of classification than the current ones under expert segmentation and equivalent in terms of prediction, demonstrating that a prediction of the natural way and not so arbitrary on the part of the experts. Another contribution that is made during these experiments is that each group can quickly identify the sales cycle that identifies it, this allows the marketing group to create prioritized actions for the books based on the sales segmentation it has and the theme, see Figure 3.33.

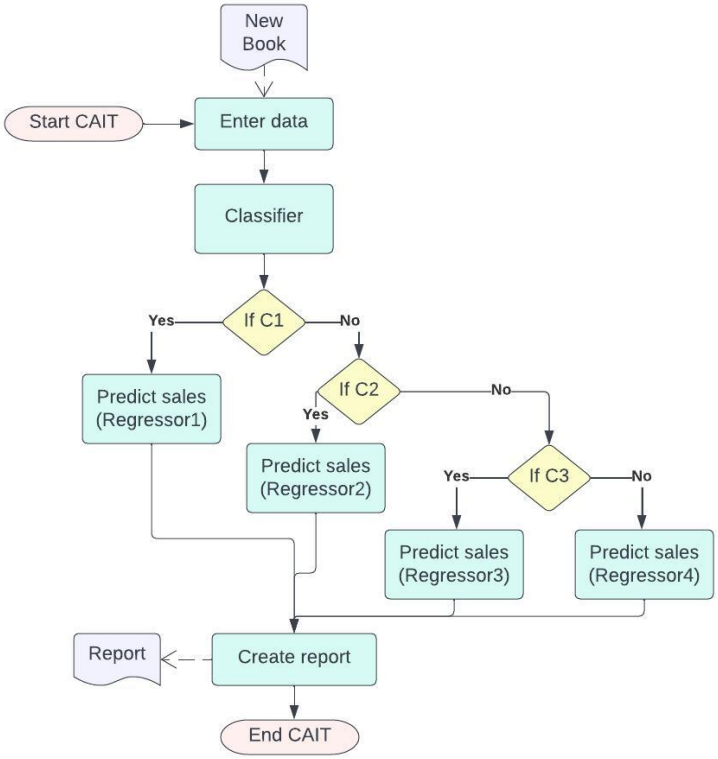

FIGURE 3.33: Diagram of the general algorithm of the CAIT tool. Source: Own elaboration.

The previous Figure allows us to understand everything previously described, visualizing the operation of the tool when a new book arrives at "La Editorial".

Chapter 4

Character personality profile

"All our experiences merge into our personality"
-Malcolm-

This chapter presents and details the second contribution made with this research. It consists of determining a psychological profile of the characters in literary works of the Fiction genre in Spanish. It consists of a tool that prevents the loss of information caused by the reduction of cacophony in the texts. The main objective is to integrate multidisciplinary professions such as Psychology, Literature and Artificial Intelligence.

4.1 Introduction

The factors that can influence the success of a book in the Fiction literary genre are completely different from those of Non-fiction. For this, it is necessary to consider fundamental aspects of fictional stories, based mainly on the narrative functions that a character fulfills, embodied in the work of Vladimir Propp [105]. Within these aspects are: a) plot, b) themes, c) structure, d) characters, e) writing, f) dialogues, g) turns, h) setting, i) genre, j) rhythm, k) emotion and l) key moments. However, this chapter will not only focus on the characters as a key piece of the work, but also proposes an analysis that allows to show the personalities of the characters.

If we analyze the dataset "Books" with which we worked in the previous Chapter, it can be observed in Figure 4.1 a greater existence of female authors than of males for the Fiction genre. Therefore, it is This chapter also makes a comparison of the profiles of characters, by gender of the author, intending to carry out a critical analysis of the work.

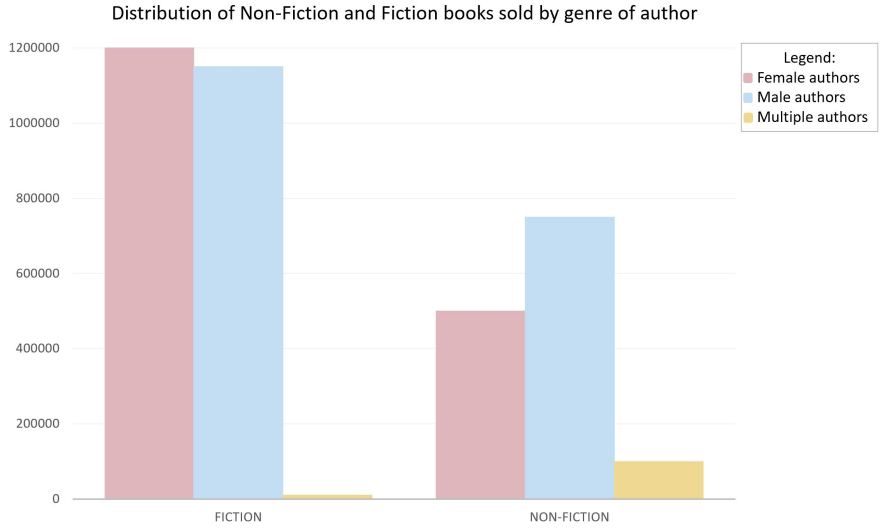

FIGURE 4.1: Distribution of Non-fiction and Fiction books sold by genre of author. Female authors predominate in the Fiction genre and male authors in the Non-fiction genre. Source: Own elaboration.

In the previous figure, it can clearly be seen that in the literary genre of Fiction, the list of sellers, although it is almost similar to sales, is more dominated by female authors and a significant value of sales of co-authors is not observed in said genre (or some authors with pseudonyms). While, in Non-fiction, a notable difference can be seen between the sales of female/male authors, the male being the outstanding gender in sales of Non-fiction books, and it is also observed that in this genre if it begins to make the presence of co-authors (or some authors with pseudonyms) more notable than in the Fiction genre.

4.2 Dataset

Based on the previous analyses, the data analysis begins with the similarity analysis between the works of the authors of both genres, focused on the Fiction literary genre. To start exploring the similarity of authors, the works of an author are first analyzed, in order to know if they have certain patterns of behavior or not Figure 4.2 analyzes "What does the writer Gabriel García Márquez have in his head?". For this, a WordCloud is proposed with the words used by the author in his works.

FIGURE 4.2: Visualizing what Gabriel García Márquez had in mind when he wrote the stories: Los funerales de Mamá Grande (Big Mama's Funerals in English), Cien años de soledad (One Hundred Years of Solitude in English), Amor en los tiempos del cólera (Love in the Time of Cholera in English). Source: Own elaboration adapted from: https://i.pinimg.com/originals/a1/48/92/a1489238bce018e368682e2f371a80ca.jpg

With this visualization it is possible to intuit a certain pattern of repetitions of keywords for the author in all these works, in this case, one of the words that Gabriel uses frequently are: house, the world, time, among others. This provides clues for further stylographic analysis of the authors of various works. For example, Figure 4.3 shows the results after the similarity of the texts between 4 authors, using the Cosine similarity which will be detailed in 5.2.3.1. Two authors have been selected by a genre and type of literary work.

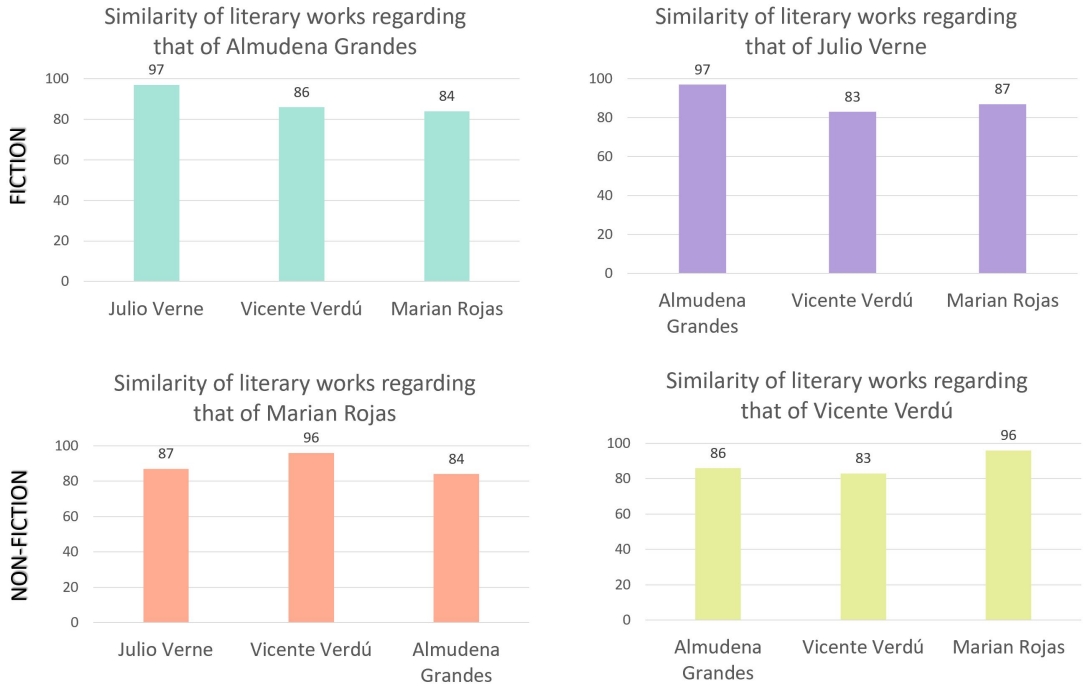

FIGURE 4.3: Visualization after the comparison of similarity of works between authors, both of the Fiction and Non-fiction genre. Source: Own elaboration.

The previous results are the contribution of a limited test for a specific work by each mentioned author (which is not applicable to all his works), but they allow to show the comparison between the authors of Fiction (Almudena Grandes and Jules Verne), turning out to be similar each other using the similarity of the Cosine, having a certain analogy since both authors write works of the Fiction genre. However, if it is compared with the authors of Non-Fiction (Vicente Verdú and Marian Rojas), a certain decrease in the resemblance is perceived, but it does not provide enough information to continue the line of character analysis after the resemblance between the authors' works. Therefore, the experiment continues, focusing directly on the extraction of the characteristics of the characters in each work.

Initially, three sets are prepared to train each model used in each of the parts of the tool: The first dataset called "NER"(10,6750 samples) has features such

as: word, part-of-speech (POS) tags, an entity tag (NER); the second dataset called "Pronouns"(1776 samples) contains the following structure: text, pronoun, position where the pronoun is found, the name of the character A, position where the character A is found, label if the pronoun really belongs to this character or not, the name of character B, the position character B is in, and again another label about whether or not the pronoun belongs to this character. For this dataset, only pronouns have been taken into account personal (he and she, both uppercase and lowercase). In addition, the cases have been defined as TRUE, FALSE, NONE (for the topics in which the pronoun does not belong to either of the two characters). Finally, a dataset called "MBTI" (8675 samples) downloaded from Kaggle [106] has been used to determine personality types according to Myers-Briggs [33], highlighting that the dataset has been translated into Spanish since no source was found in that language. The observations (texts written by people) contained in this dataset have been labeled by the psychologist for the Myers-Briggs personality type.

In Table 4.1 the datasets used for the development of the research are shown. They are divided into train, validation and test sets to avoid overfitting of the models.

TABLE 4.1: Dataset information: the Corpus column shows the name of each of the datasets, the Sets column shows the different partitions that are performed for each of the datasets and the Nb Data column shows the sample size for each of the partitioned datasets. Source: Own elaboration.

ID	Corpus	Sets	Nb Data
1	Books	Train/Test (70/30)	818/351
2	NER	Train/Test (80/20)	85400/21350
3	Pronouns	Train/Val/Test (46/44/10)	816/781/179
4	MBTI	Train/Val/Test (60/20/20)	5205/1735/1735

4.3 Proposed tool

This section describes the proposal for a new support tool resulting from the study [107], so as not to lose any type of information that can later recreate a character profile in a literary work.

Figure 4.4 shows the initial phase of this study. It consists of collecting data to create a corpus focused on literature and in the Spanish language. The data described above in the Table 4.1 are collected from multiple sources, such as: ciudadseva.com, Wikipedia and Kaggle and goes through a cleansing phase. Noisy or incorrect data are removed. The common normalization in Spanish is the elimination of accents (except "él", *he, in English*), which is required for the detection of the personal pronoun.

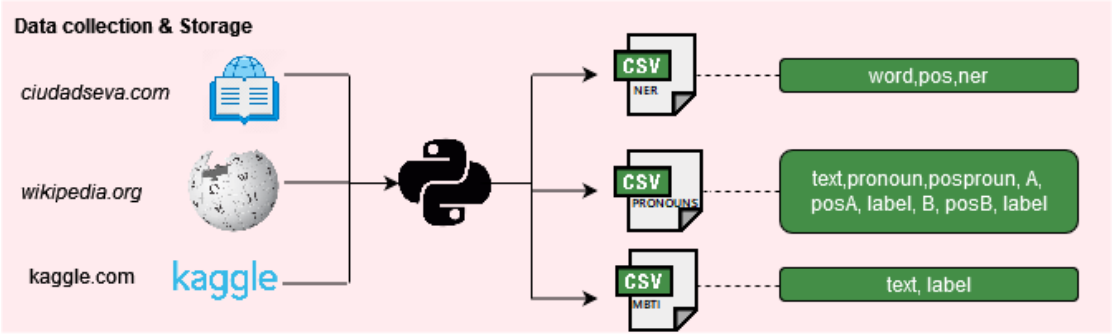

FIGURE 4.4: Phases for the creation of a support tool to determine the profile of a character. Data collection phase: collect the data from different sources. Then they go through a preprocessing that consists of a cleaning aimed at eliminating noisy or incorrect data, the common normalization in Spanish is the elimination of accents (except the personal pronoun "él", *he, in English*). Finally, the training/test dataset is separated for its later use in the model. Source: Own elaboration.

Subsequently, Figure 4.5 continues with the phase of the proposed tool, where its composition and logical order of processing are shown: the input data go through the first model "Character recognition", then they enter the "Pronouns of belonging" to determine those sentences that belong to the characters given the personal pronoun used and with this add new characteristics to the characters and finally, once the texts belonging to the characters have been selected, they are entered in the "Character profile" model, which will indicate the profile of the character.

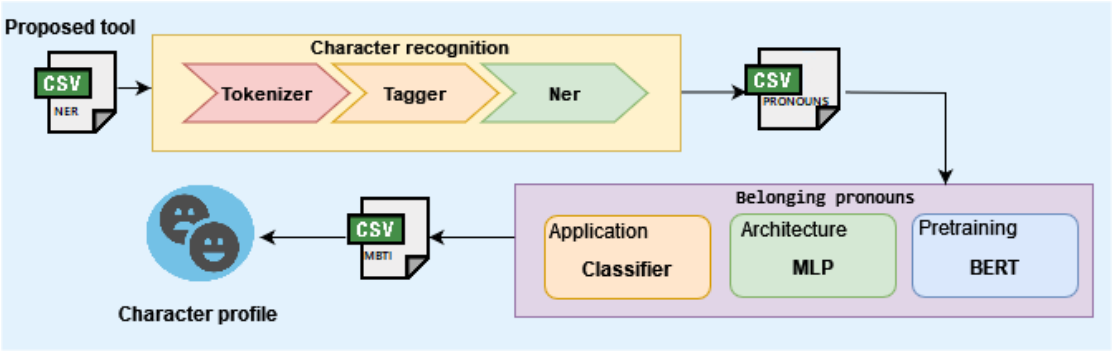

FIGURE 4.5: Phases for the creation of a support tool to determine the profile of a character. Proposed tool Phase: once the model has been trained, the logical order of processing new data is shown. Initially, the characters and the features that define them are found, then new features are added that can be obtained from the personal pronouns that correspond to each character and finally, a sentiment analysis is carried out to create a character profile. Source: Own elaboration.

Finally, Figure 4.6 shows the validation and visualization phase, where the proposed tool is checked, and the results are displayed. During the validation phase, internal validations are carried out on the models. And in the visualization phase, the data is presented in an illustrated or graphical format, in a way that allows decision makers to see the analyzes presented visually.

FIGURE 4.6: Phases for the creation of a support tool to determine the profile of a character. Validation Phase: find the internal validations that are performed on the models and finally the data is presented in an illustrated or graphical format, in a way that allows decision makers to see the analysis presented visually and, therefore, can grasp difficult concepts or identify new patterns. Source: Own elaboration.

4.4 Experimentation algorithms

The algorithms used in each of the models that make up the tool are described below.

4.4.1 Part A: Character recognition

For this first part of the tool, the spaCy [108] library is used, because it is currently the fastest NLP Framework and has one highly optimized tool for each task, in this case, named entity recognition. For a better understanding of the training algorithm used by this tool, see the following steps:

1. Initialize the model weights randomly.
2. Predict some examples with the current weights.
3. Compare prediction with true labels.
4. Calculate how to change the weights to improve the predictions.
5. Update weights slightly.
6. Go back to step number 2.

4.4.2 Part B: Belonging pronouns

Based on one of the related works mentioned, in this second part, the text is subdivided into a range of 300 words to search the text for the characters extracted in the previous phase. This will allow creating a dataset that goes through the second part of the "Membership Pronouns" tool, intending to add (if they exist) other character features and thus avoiding the loss of information by reducing the cacophony.

For training, Bidirectional Transformer Encoder (BERT) representations are used as presented by Devlin [109], a state-of-the-art model that uses self-attention layers to understand the meaning of the opposite sides of each word. As architecture, a multilayer perceptron (MLP) is used [110, 111]; the typical architecture of this type of network is made up of several layers of nodes with complete interconnection between them.

The following will describe how this model has been trained with the Wikipedia example dataset:

1. Step 1: Input a batch of mini-inputs with N random samples from the training dataset, previously labeled (TRUE/FALSE/NONE, which is a real output).

2. Step 2: After the relevant calculations in each layer of the network, the output predictions are obtained (forward propagation). The hidden layers have a Rectified Linear Unit (RelU) as their activation function.

3. Step 3: The cost function/loss function for said minilot is evaluated, with a cross-entropy function (Eq 4.1), since a classification of whether or not the pronoun belongs to the characters is desired.

$$H(p,q) = -\sum_{x} p(x) \log q(x) \tag{4.1}$$

where p: is the real sample distribution, q is the model to be estimated. Both follow a 0-1 distribution;

4.4.3 Part C: Personality Profile

In this last part of the tool, advanced techniques such as Transfer Learning are used. As many can guess, the growth of Deep Learning is increasing, and they often have to manage long computation times and many resources. Therefore, to begin this phase, a previously trained model is used as a starting point. In this case, Transfer Learning allows to quickly develop efficient models and solve complex Computer Vision or NLP problems (as this study shows). Figure 4.7 shows a graphic example of how this technique works. Given a domain A, it is entered into a model that will be divided into two layers, Body/Head. The head is the specific task that you want to attack, and the body, where the weights are stored, which gives an idea of how the language works. Statistical correlations are also stored here. If this model is trained, it is possible to have another different domain, where these weights from model A can be used and used to initialize model B, only the specific task would have to be adjusted in this second model.

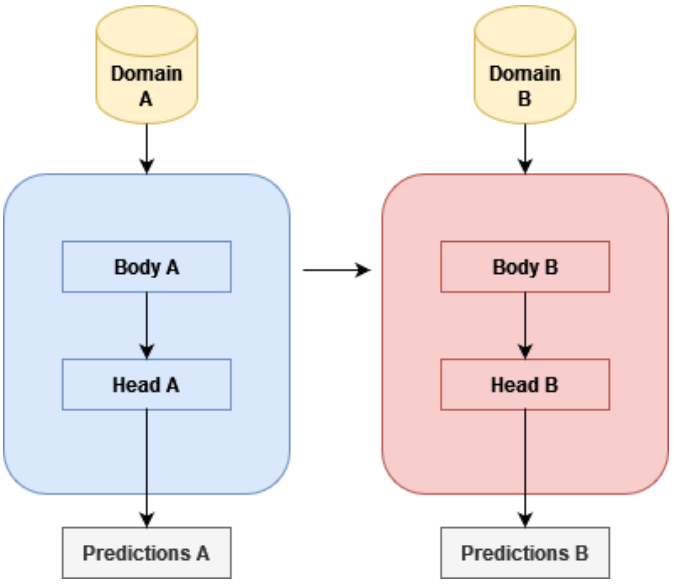

FIGURE 4.7: Transfer learning from one Domain A (global) to another Domain B (specific to a new task). Source: Own elaboration.

Therefore, using this technique, the weights are loaded into a previously trained transformer model, specifically for the fill and mask task. These base tasks allow adjusting the models to classification tasks or answer questions. This study focuses on classifying the 16 personality types, differentiated into 4 axes: [Introversion (I) – Extroversion (E); Intuition (N) – Sensation (S);Thinking (T) – Feeling (F) ; Judge (J) – Perceive (P)] according to the MBTI created between 1940-1950 [33, 112].

In Table 4.2, we show the 16 personality types, for more details of their meaning, visit the official page of the Myers-Briggs Foundation.

TABLE 4.2: The 16 personality types, according to the Myers-Briggs Indicator, based on 4 different axes: [Introversion (I) – Extroversion (E); Intuition (N) – Sensation (S); Thinking (T) – Feeling (F) ; Judge (J) – Perceive (P)]. Source: Own elaboration.

ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTP	ESFP	ENFP	ENTP
ESTJ	ESFJ	ENFJ	ENTJ

4.5 Results

SpaCy has a pre-trained model in Spanish based on a news dataset to find PERSON labels, however, the classifications do not fully fit the training dataset that has been obtained. For a better understanding, in Figure 4.8 (a), where the work "The prodigious afternoon of Baltazar" by Gabriel García Márquez is selected as an example. To start training the proposed model, we start with the configuration of the pipeline, which will be adapted to be in Spanish and the lines with which it will work, in this case the following are added: Tokenizer, to divide the text into tokens (or words); Tagger, to be able to learn the labels of each of the parts of speech, with which you will learn what word is an article, a noun, an adjective, etc.; NER, so you can learn the labels of when a token or word is a person (PER), a place (GPE), a building (FAC), etc., but with the format `IOB`: `B`, which indicates where the noun phrase begins; `I`, describes that the word is inside the current noun phrase and `O`, end of sentence. This research will only focus on training the model to determine with better precision which token or word is a person (taking it into the context of this thesis, a character), since the labeling for the training set has been done manually. Once the model is trained on the Literary Text-Based Approach dataset, it can be seen that it fits a bit better than a pre-trained model. Figure 4.8 (b) shows this adjustment, making a comparison with the same text in Figure 4.8 (a). More details, see Annex B.0.1.

Figure 4.9 shows the NER distribution (especially that of characters) by both the pretrained model and the fitted model. As can be seen, at the level of recognition, the characters found fit more with the model fitted with a dataset in the literary context than with a pre-trained model. In the Figure only the first 4 books tested are shown, for more details see Annex B.0.2.

Once the characters have been identified, the importance they have in the work can be extracted, given the number of repetitions within the text. This allows to give another type of contribution to the work and is the determination of whether the work is choral or not. A work is choral when several characters are connected with the work as a whole and all with the same importance as the protagonist. Figure 4.10 shows the importance of the characters indicates that most of the works are not choral, since not all of them have the same level of interest for the author in each story.

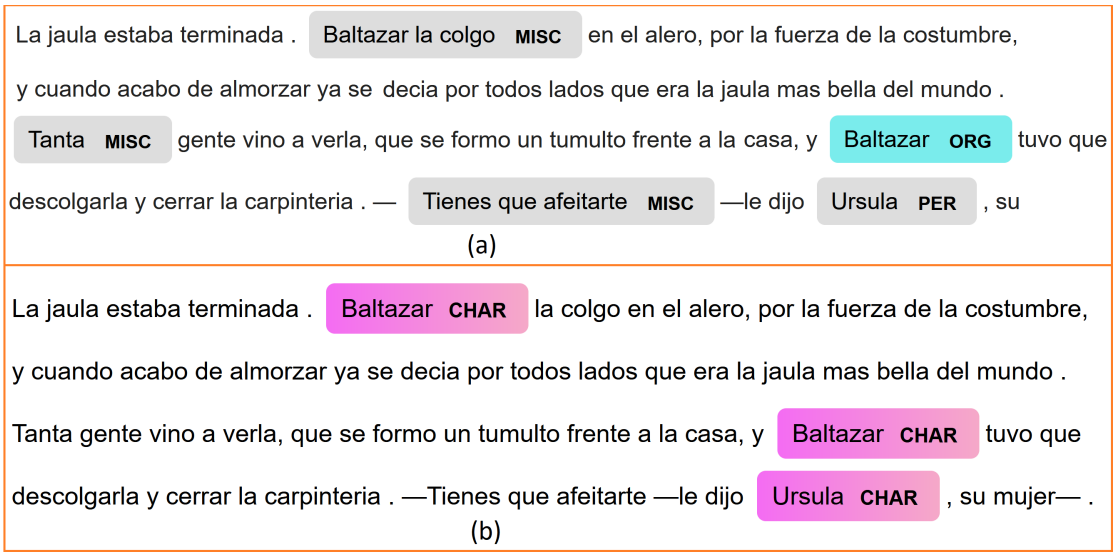

FIGURE 4.8: (a) Labeling result with the prediction of the pretrained model of spaCy, being: PER-person, MISC-miscellaneous, ORG-organization. Only a PER type tag is correct, as can be seen. (b) Labeling result with the prediction of the model trained with the literature dataset, where: CHAR-character. Both texts (a and b) in English mean: *The cage was finished. Baltazar hung it on the eaves, by force of habit, and when he finished lunch everyone was already saying that it was the most beautiful cage in the world. So many people came to see it that a riot broke out in front of the house, and Baltazar had to take it down and close the carpentry. - You have to shave - Úrsula, his wife, told him.*. Source: Own elaboration.

Taking as a reference the related work that guaranteed the relationship between two characters in a range of 300 words [57], the first experiments are demonstrating this fact. With the characters identified in the first part of the proposed tool, the texts are then separated under the range established by the study. In Figure 4.11 one of the stories of the famous writer Gabriel García Márquez is taken as a reference, where the relationship between each of the characters in the play "Love in the time of cholera" is observed. Each node of the graph represents the character in question; the size means the number of repetitions that the character has in the work. The vertices indicate the relationship of each of the nodes.

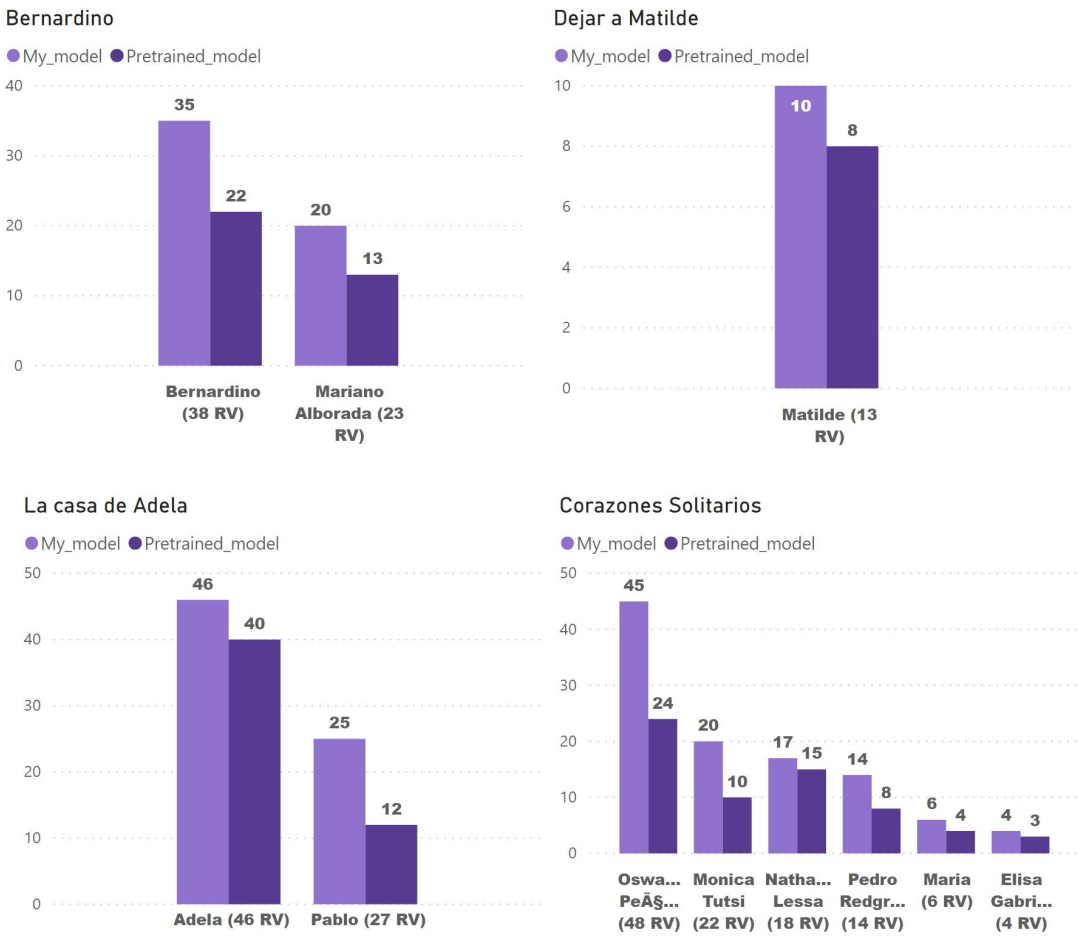

FIGURE 4.9: Result of characters found with the pretrained model of spaCy versus model fitted with the literary theme. The real number of appearances of character in the work on the X axis next to the name of character. Making a comparison between the models, it can be seen that the results of the adjusted model are much more accurate than those of the pretrained model. Source: Own elaboration.

For the training of the Part B of the proposed tool, the hyperparameters of the neural network are adjusted (learning rate and weight decay) to obtain more optimized results, as shown in Table 4.3:

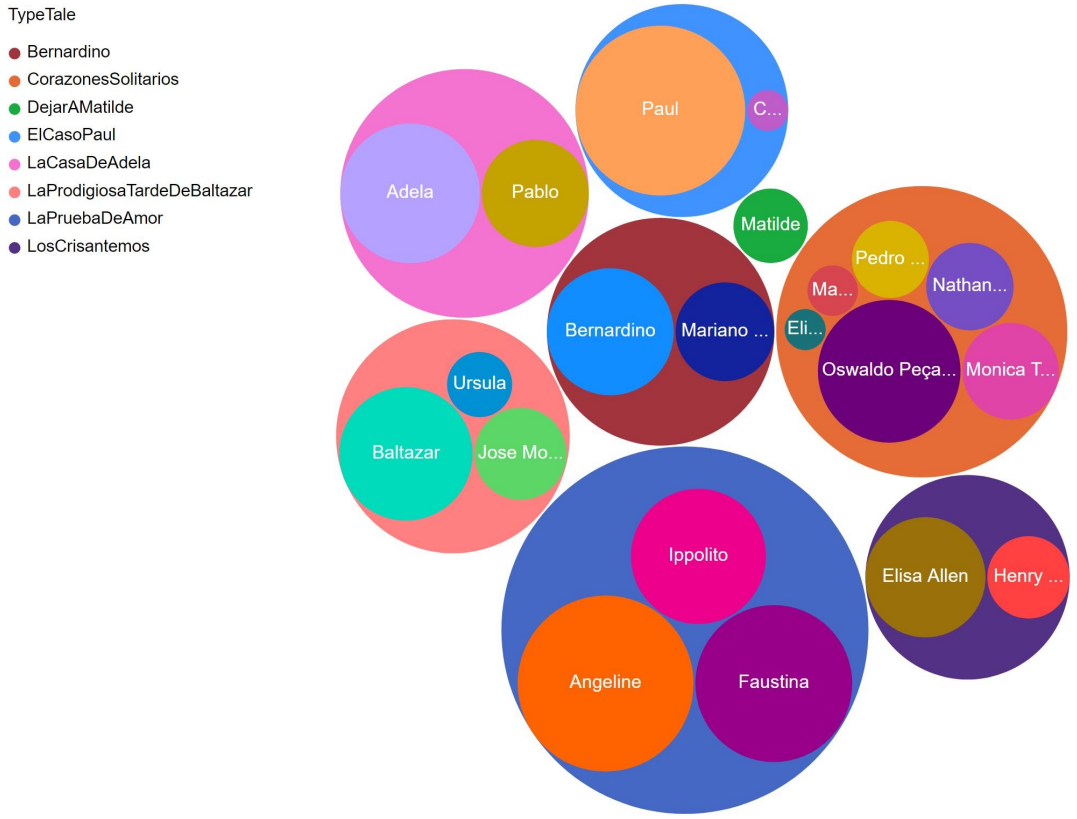

FIGURE 4.10: Visualization of the importance of the characters for each story. The characters are grouped according to the story to which they belong. It is evident that, given the repetition of the character in the work, some are mentioned more than others by the author, indicating that the works are not choral. Source: Own elaboration.

TABLE 4.3: Result of training pronouns dataset. The best results are offered by the v4 model with a learning rate and weight decay of 0.005. Source: Own elaboration.

Model	Learning rate	Weight decay	Accuracy(%)	Precision(%)
v1	1e-3	5e-3	75.99	50.24
v2	1e-2	5e-3	75.99	50.13
v3	5e-4	5e-3	76.32	56.66
v4	**5e-3**	**5e-3**	**78.29**	**80.19**
v5	5e-2	5e-3	73.03	37.45
v6	5-e3	0	71.38	48.2
v7	5-e3	5e-4	68.75	39.35

Once the model has been trained, it will be able to identify which character the pronoun of the next reference made by the author in the work belongs to, in order to avoid cacophony. In order to understand what the model does, in Figure 4.12 the prediction made with this second part of the proposed tool is shown.

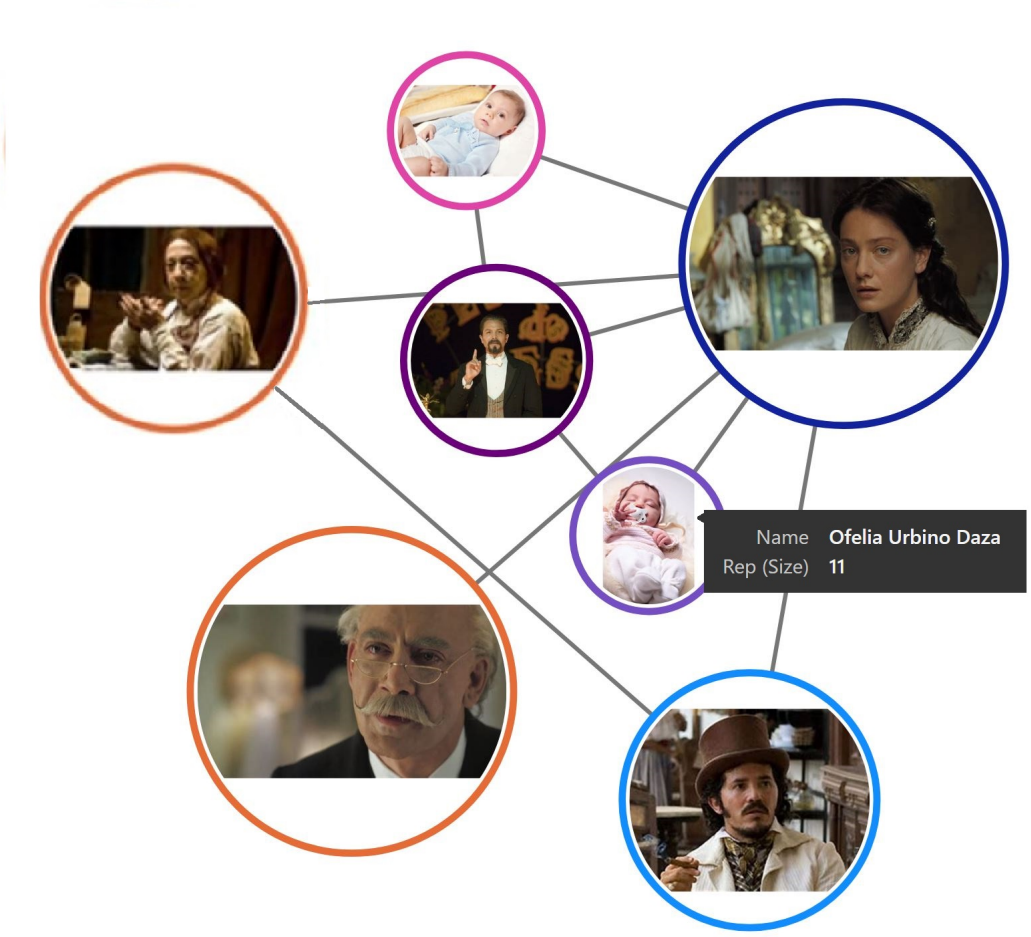

FIGURE 4.11: Visualization of the relationships between the characters of the play "Love in the Time of Cholera". Source: Own elaboration adapted from: http://tse1.mm.bing.net/th?q=Lorenzo%20Daza,http://tse1.mm.bing.net/th?q=escol%C3%A1stica%20daza, http://tse1.mm.bing.net/th?q=Fermina%20Daza, http://tse1.mm.bing.net/th?q=Juvenal%20Urbino, http://tse1.mm.bing.net/th?q=Florentino%20Ariza,http://tse1.mm.bing.net/th?q=ni%C3%B1o%20bebe and http://tse1.mm.bing.net/th?q=ni%C3%B1a%20bebe

Important, the example will only be in Spanish, since that is how the model is trained, and therefore it cannot be viewed with another example.

Carlos es muy bueno, así como Juan. Aunque él es mucho más estudioso.

	ID	Text	Pronouns	A	B	Prediction
0	0	Carlos es muy bueno, así como Juan. Aunque él ...	él	Carlos	Juan	A

FIGURE 4.12: Visualization of the prediction made with Part B of the proposed tool, where the person to whom the personal name used belongs is identified. Only in Spanish, since the model is trained that way. Source: Own elaboration.

For training of the Part C of the proposed tool, a model generated by transfer learning is used. A base model, Roberta (Robustly Optimized BERT Approach) [113], is used with this model. It is a slightly modified BERT architecture with more effective pretraining. This model is trained with 160GB (Wikipedia, Common Crawl, News, Book Corpus, Webtext Corpus) with NVIDIA Tesla V100 for 24h, at a cost of $72000. This makes it impossible to increase these costs, hence the reason for using the Transfer Learning technique.

This model is adjusted by different hyperparameters until finding the one that best fits the MBTI dataset used, reaching an accuracy of 54%. Specifically, the optimal hyperparameters are found in Table 4.4. Although other evaluation metrics such as recall and F1 could have been used, only Accuracy is used, with a view to verifying similarities or not with the base study [58].

TABLE 4.4: Hyperparameters used in personality model.

Hyperparameters	**Value**
Learning rate	1e-05
Train batch size	16
Eval batch siz	16
Seed	42
Optimize	Adam with: - betas = (0.9,0.999) - epsilon = e-08
Lr scheduler type	linear
Num epochs	20

As it is well known, the problem of "solving AI" will not be solved by a single company or research institution, but by a culture of sharing knowledge and resources. That is why the Hugging Face Hub platform is used where you can share and explore Machine Learning models, datasets and demos; and as much as the base model of Roberta is used from this platform, this research contributes to this part of the model for the determination of the personality of a character in the Spanish language. The model is available on HuggingFace [114].

Finally, creating the personality profile in Figure 4.13 details in an easier and more visible way the content that is shown at a global level in Figure 4.14, where the analysis of the eight stories will be shown. In all the spider graphs, it will be possible to visualize the graphic representation of where the personality of each of the characters in the stories is found. Each attribute/personality shown are those associated with the 16 Myers-Briggs Personality Types.

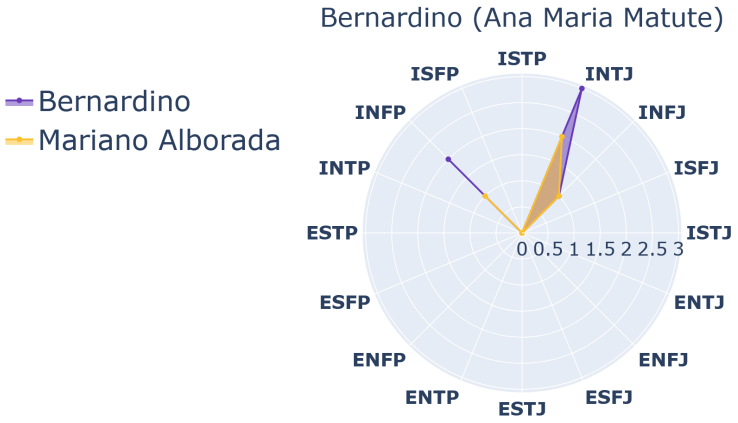

FIGURE 4.13: Radar or spider graph of the set of personality types plotted on axes from the same point. The scale represented in the spider graph responds to the frequency of this type of personality in the texts belonging to the character entered into the model. Source: Own elaboration.

Figure 4.14 you can clearly see the characteristics of each of the characters in the treated literary work. The female characters treated by male authors are described more among the personality traits [INTJ-ISTJ], that is, they are treated as having original minds, independent, calm, reliable and loyal. While the female characters treated by female authors are also described as having an INFP personality, that is, idealistic, loyal to their values, adaptable, flexible and tolerant unless a value is threatened.

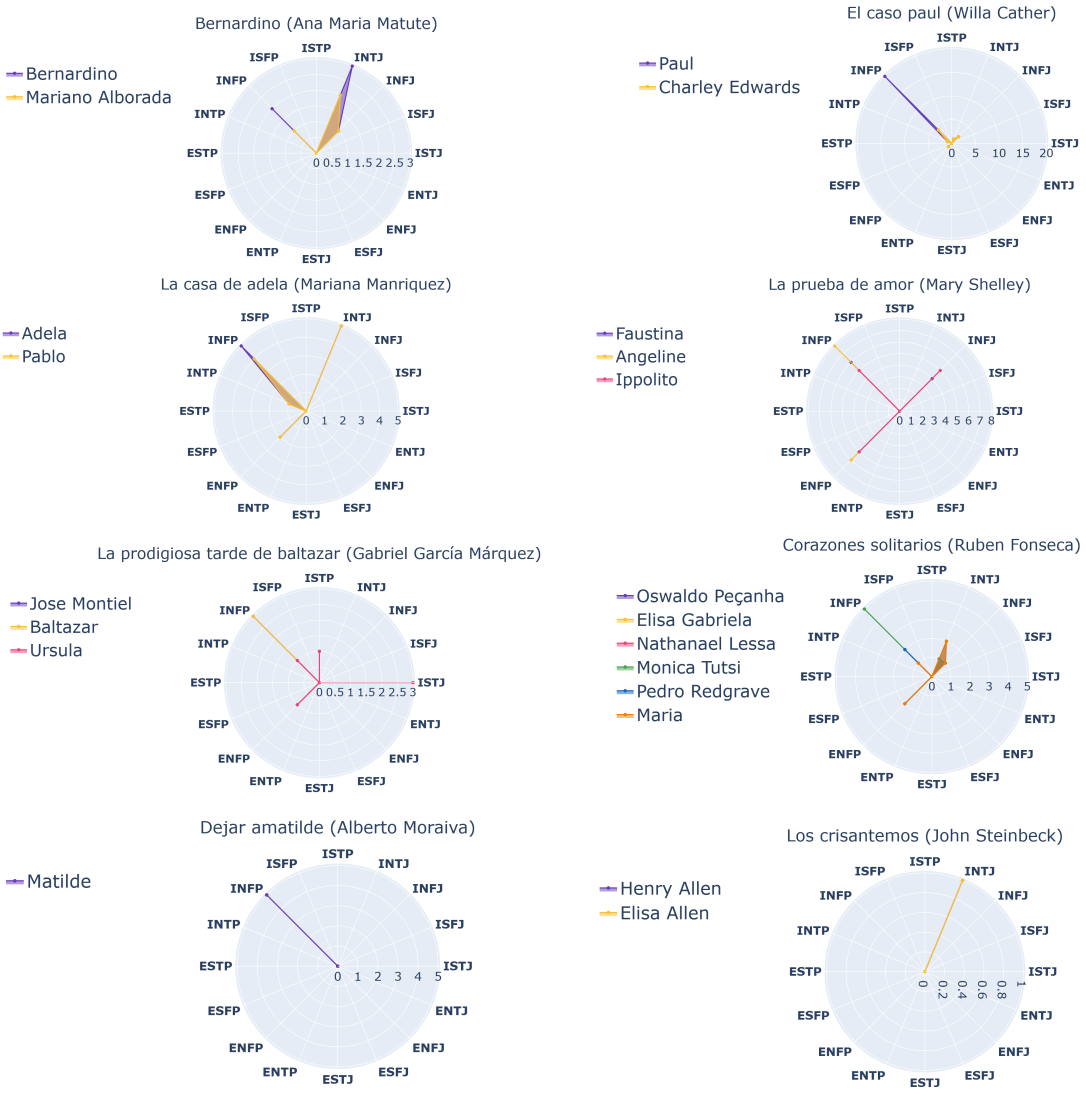

FIGURE 4.14: Result of the personality of each character in the stories. It is observed that for both female and male authors, the male characters present the same type of personality, however the female characters of the authors are more independent and for the female authors they are more flexible and adaptable, based on the Myers-Briggs personality descriptions. Source: Own elaboration.

In the case of the male characters treated by the authors, they have personality traits [INFP-INTJ-ENFP], therefore they have a warmly enthusiastic and imaginative personality, they are spontaneous and flexible, they tend to trust their capacity for improvisation and their verbal fluency, original and loyal.

4.6 Summary

This section not only contributes to the main task, which was the creation of a psychological profile, with all the information about the character (both explicit through the name and relative through the personal pronoun); but also with the advanced techniques of Artificial Intelligence in the world of NLP, it can be determined if a work is choral or not.

Chapter 5

Recommender system

"Don't look for customers for your products, look for products for your customers"
-Seth Godin-

The third contribution made in this thesis is detailed in this chapter. This is focused on the personalized recommendation of books, beyond the sales history. The main objective is to stimulate the reader's empathy with the book, so that reading is encouraged.

5.1 Introduction

Recommender systems are part of an information filtering system. In the publishing sector, one can often find book recommendations based on sales history (for example, Amazon) or even recommendations based on the arguments of literary critics. Both responding in many cases to the interests of the publishers and not of the readers. This last contribution to the doctoral thesis project will be based on the writing of the readers (in this case, a Twitter user). Next, the system created will be detailed. This recommends Fiction books, since they are the best-selling books worldwide, as can be seen in Figure 5.1 and Non-fiction Books, since they are focused on everyday reality, as can be seen in Figure 5.2.

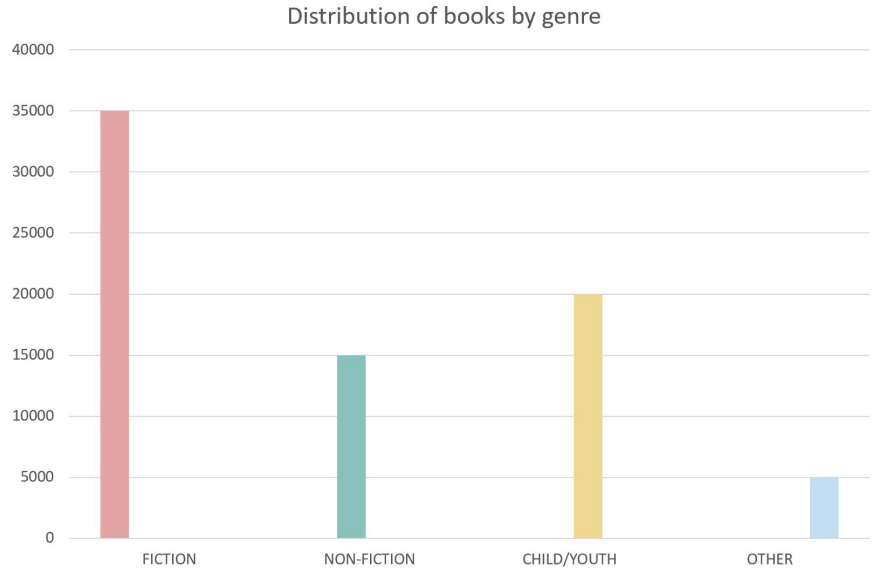

FIGURE 5.1: Distribution of book sales by literary genre. Where the best-selling genre is Fiction. Source: Own elaboration.

FIGURE 5.2: Wordcloud with the most words written on Twitter. Source: Own elaboration.

In the previous figure, the most common words in the Twitter community are shown, which allows us to intuit the most frequent topics such as: video games, employment, politics, sports, wellness.

5.2 Fiction books

The entire process of creating the first part of the recommender based on Fiction books will be detailed below. Explaining the dataset to be used, the proposed tool, the experiments and the results obtained from them.

5.2.1 Dataset

To obtain the reader's way of expressing themselves, data have been collected from the social network Twitter, the second social network with the highest number of users and the easiest to access its data through its own API [115], and they have been saved in a MongoDB database [116]. It has around 2,590,705 samples from the period from April to May, 2022

Initially, the system is fed with a dataset resulting from the previous contribution [107] (150 samples) containing the text belonging to each character in the book, the title of the book it belongs to, and the personality associated with the character. Subsequently, these cases will be expanded with new cases once the Review phase is over. The content of the new cases will be the same: written text, the book to which it is associated, and the personality associated with that type of writing.

5.2.2 Proposed tool

The publication treatment of Non-fiction and Fiction books is totally different from each other, as has been mentioned throughout the thesis. Therefore, this same decision has been maintained for the recommendation of books. On the one hand, the recommendation of the Fiction books will be based on the similarity of the texts (both the tweet and the book) and the Non-fiction books will be based on the subject they deal with. This section describes the system called, Artificial Intelligence Based Reasoning with Natural Language (BRAIN L), for recommending Fiction books based on readers' own writing, resulting from the study [117]. The system begins with the entry of a problem (in this case the text of the tweet in search of a book recommendation). Subsequently, it goes to the *Retrieve phase*, which consists of searching for potential cases to solve this problem, searching within the Memory Cases for the books with texts most similar to the problem. Once evaluated which is the optimal solution of the multiple cases, in the *Reuse phase*, the proposal is provided to the user. The quality of said solution is also evaluated, during the *Review phase* by the experts for its subsequent storage

as a new memory case, in the final phase, which is *Retain*. The general design of the system is shown in Figure 5.3.

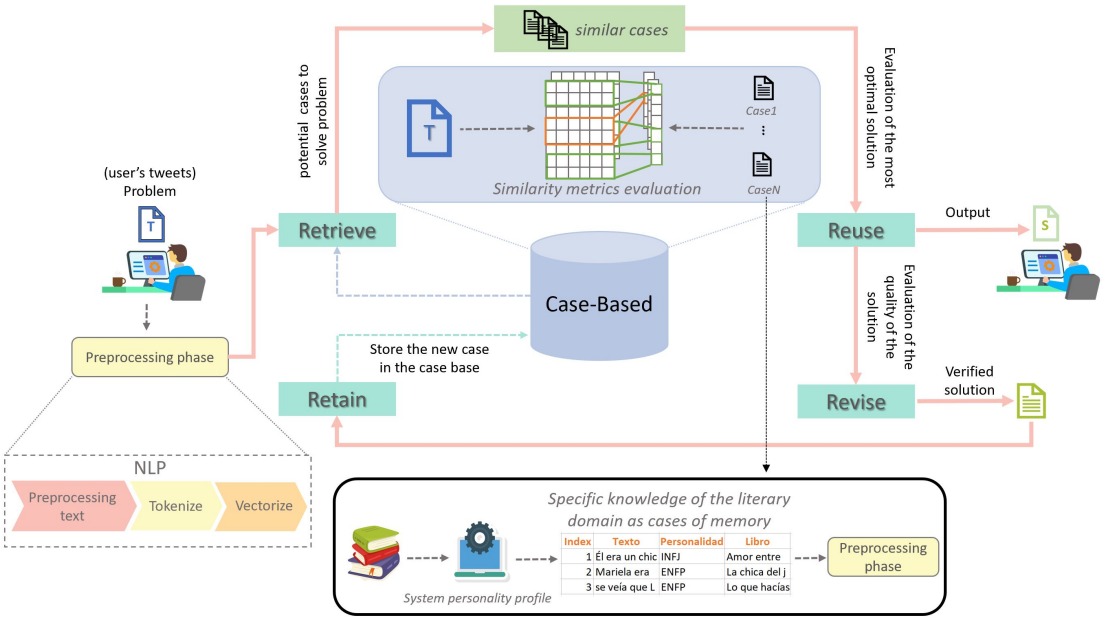

FIGURE 5.3: Case-based system architecture combined with NLP for Fiction books. Based on how a reader expresses himself, similar cases are searched in the database. For this, NLP techniques are used for text processing and similarity evaluation metrics between them. Finally, the best case is returned. This case can be reused if an expert validates it and considers it appropriate to retain in the case memory. Source: Own elaboration.

5.2.3 Experimentation algorithms

The experiments carried out on each of the genres of books treated are described below. The techniques used to experiment with each genre are distinguished, after the separation of each one.

5.2.3.1 Retrieve phase

During this phase, the text expressed by the reader (Twitter user) has been entered. The text is first processed, initially grouping all the tweets by username, thus ensuring that there is valid content for later analysis. Those characters that can cause noise in the model, such as hashtags, urls, stopwords, among others, are eliminated. Once the text is clean, tokenization is carried out not only for the input text, but also for memory cases. The tokenization process allows a text string to be encoded into human-readable token ID integers. Subsequently, the texts are vectorized, since it is the optimal way to determine the similarity. There are different metrics for evaluating the similarity of texts, depending on the grouping of the texts and the embeddings performed. The metrics to use are:

- Cosine: It is a measure of the similarity between two vectors in a space that has a product in its interior, with which the value of the cosine of the angle included between them is evaluated. Figure 5.4 shows the calculation of the metric for a better understanding of the concept. This calculation uses Eq. 5.1.

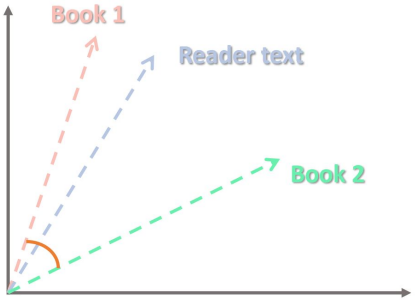

FIGURE 5.4: Calculation of cosine similarities between two books and a text entered by the reader (Twitter's user). Source: Own elaboration.

$$Cosine(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.1}$$

where A and B are the texts to compare.

- Jaccard: The Jaccard index measures the degree of similarity between two sets, regardless of the types of elements, since it is based on the intersection of the sets. Figure 5.5 shows the calculation of the metric for a better understanding of the concept. For the calculation, Eq 5.2 is used.

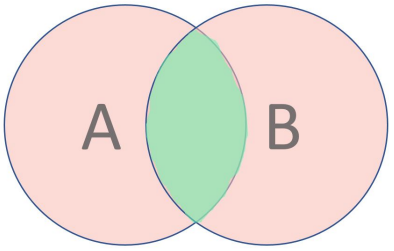

FIGURE 5.5: Venn diagram of the two texts to be analyzed. Where A is the text entered by the reader and B the text to be compared from the book. And the similarity is calculated by dividing the intersection (green color) and the union of the texts (pink color). Source: Own elaboration.

$$Jaccard(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{5.2}$$

where A and B are the texts to compare.

- SoftCosine: This measure of similarity starts from the same basis as cosine similarity, but generalizes it further by considering similarities between pairs of features [118]. This calculation uses Eq 5.3.

$$Softcosine(a, b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}} \tag{5.3}$$

where the matrix s_{ij} represents the similarity between features.

Regarding the embeddings, these are:

- Word2Vec: Embeddings are performed using a neural network model to learn word associations from a large corpus of text. For this reason, it is good for semantic analysis. [119]
- GloVe: It is a type of word embedding that encodes the co-occurrence probability relationship between two words as vector differences. [16]
- BERT: It is the most advanced technique in the NLP world. It allows extracting characteristics of the embeddings made to the text, useful for semantic information. And furthermore, it adds a special MASK token

that allows it to make predictions, finding the words that provide relevant information to the text. [20]

Once the potential cases to solve the problem have been determined, in this case, the most similar books according to the reader's writing, the next phase is continued.

5.2.3.2 Reuse phase

This phase performs an evaluation of the optimal solution and returns it as the proposed solution for the new problem. It is likely that the retrieved case, if similar enough, is likely to contain a suitable solution. To do this, the threshold is set at 50%. This allows that once the similar cases have been recovered and given that they present a score; it is analyzed whether or not the highest value exceeds this threshold. In cases where it is not passed, 2 possible books are recommended (from highest to lowest score) with a recommendation reliability message "Recommendation reliability: -50%". While, if it passes it, the case with the best score is displayed and the message "Reliability of the recommendation: +50%". If the case exceeds the threshold, it indicates that it is apt to be stored as a possible case within the base case, thus moving on to the next phase.

5.2.3.3 Revise phase

A CBR agent usually requires some feedback to know what is going well and what is going wrong. It is usually done by simulation or by asking a human oracle. For this reason, during this phase, policies are established in the system so that experts in the field can validate whether or not the new case is appropriate to be included in the base case.

The policies to be established:

- If the case is validated by a single expert, the case will not be inserted.
- If the case is validated by two experts, the criterion of not inserting the case must be justified.
- If the case is validated by three experts, it will be inserted directly into the database.

5.2.3.4 Retain

If the case solution generated during the review phase needs to be retained for the future problem-solving, the base case is updated with a new case learned in the retention phase. The data that is stored is: the text of the tweet, the recommended book based on the wording of the tweet and the personality associated with the recommended book. In this particular case, there is no implication in storing this data, since it contains the same fields.

Once the design of the system has been understood, and for a better understanding of the system, the behavior of the entire CBR cycle using NLP will be illustrated in the next section with a use case.

5.2.3.5 Case study

Before showing the results and analyzing them, a case study is proposed, where the complete operation of the CBR is visualized with a real case. This helps to understand the experiments in the following section. Figures 5.6, 5.7, 5.8 and 5.9 show the internal flow of each of the phases of the proposed system.

```
                    BRAIN L - Expert system
New case: No hace frío, pero me encuentro envuelto en una frazada de Tramas
´

Retrieve phase:
Similar Books:
{0: 0.10416666666666667, 1: 0.029411764705882353, 2: 0.1388888888888889, 3: 0.7857142857142857, 4: 0.13888
```

FIGURE 5.6: Retrieve phase. The similar cases found in the base case are shown, it contains the index of the book that is similar and the score of the similarity calculation. Source: Own elaboration.

The previous image shows how a new case is introduced, specifically, a tweet. When searching for similar cases, internally, the system will return the most similar books and the similarity score.

```
Reuse phase:
Most similar is book with index:  3
Recommendation reliability: +50%

Answer: The book ElPerseguidor  is recommended. The reader has a personality type right now: ISTP
```

FIGURE 5.7: Reuse phase. Obtaining the highest score. It is analyzed if it exceeds 50% reliability that the recommendation is correct. If you exceed it, that is the case. The reader is shown the recommended book, including the currently associated personality type, based on its wording. Source: Own elaboration.

Having several similar cases, internally the system, from a threshold established in this case, 0.50 assesses which of the scores is the one that exceeds said threshold. This will create an internal validation of the predicted recommendation, before being shown to the end user.

```
Revise phase:
Enter validation (expert 1):y
Enter validation (expert 2):y
Enter validation (expert 3):y
The new case is suitable for storage!
```

FIGURE 5.8: Revise phase. During this phase, the policies are applied so that the experts with the validated solution determine whether or not it should be stored in the memory cases. Source: Own elaboration.

If the case selected as optimal and shown to the user can serve as an experience for future searches in the system, certain policies are established. The same if they are shown on the screen to the experts, for their subsequent approval of the insertion of the new case in the memory base.

```
Retain phase:

The size of the current case base is 150

The new case to be saved contains the following fields:
Book:  ElPerseguidor
Text:  No hace frío, pero me encuentro envuelto en una frazada de Tramas
Personality:  ISTP

The case base data has been updated, now the size is 151
```

FIGURE 5.9: Retain phase. During this phase, the new validated case is stored. Source: Own elaboration.

This case study complies with all the previous phases, and after the unanimous approval of the experts, it is stored within the base case.

5.2.4 Results

Once the operation of the system with the case study has been visualized, the results of the experiments carried out are presented.

5.2.4.1 Dataset

Tweets

As this research is carried out on the Spanish literary domain, the data is limited only to the national territory of Spain. Duplicate data is removed. Only those tweets with more than 20 words are used, thus ensuring that there is valid content for later analysis.

Tweet texts are in the form of unstructured data and are full of noise and unwanted information. For example: special and repeated characters, unwanted spaces, emojis, a URL for both: a video, news, etc. and hashtags. That is why before using these data, a previous data preprocessing step is required. The abstract idea of this preprocessing can be seen in Figure 5.10.

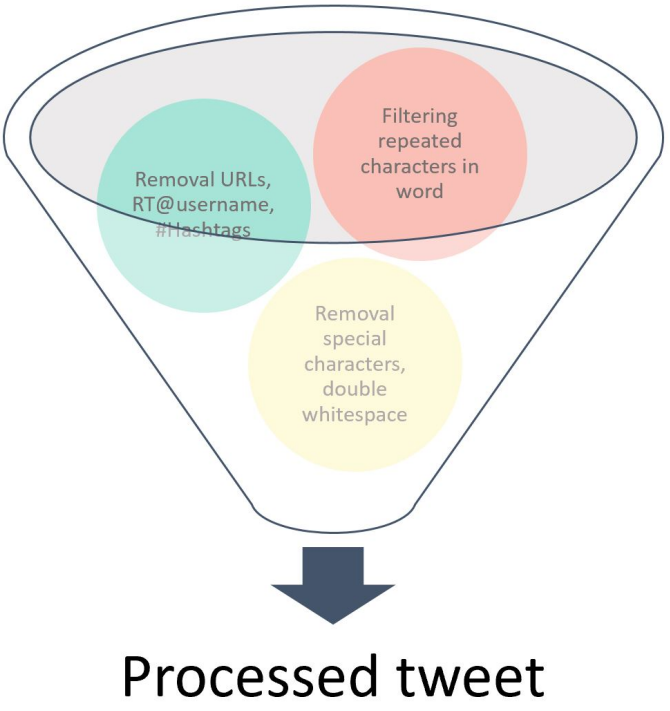

FIGURE 5.10: Filtering policy applied to the texts of the tweets. Source: Own elaboration.

Finally, the tweet is preprocessed, it is grouped by a user, keeping a total of 1500 samples.

Fiction books

Memory cases were already treated for use in the old system, which returns, given the character's text, the personality type. To store these cases, make it easier for them to be shared from different readers at the same time, avoid redundancies and improve the organization of our system, a MongoDB database has been used.

5.2.4.2 Retrieve phase

For the training phase of the system, fragments of untrained books are taken as test data, in order to verify the effectiveness of the aforementioned techniques. The modification of the words is also tested, but maintaining the same meaning of the sentence, to test the semantic analysis of the techniques. This is what is labeled "Different Text" and is the best way to determine if the system will be able to recommend books correctly.
During this phase, the different configurations mentioned in the previous section are tested to find the optimal for our recommendation system. It is important to mention that cosine similarity calculations require embedding in the text; while in Jaccard it is not necessary since it works with the number of elements per sets. The configurations made for the calculation of the similarity with Cosine are described below.

For the **Word2Vec embedding**, the embedding of the spaCy library by [120] itself is used with an `es core news lg model` and the Cosine similarity is calculated. The results are shown in Table 5.1.

TABLE 5.1: Cosine Similarity Calculation results with Word2Vec Embedding.

Type of text	Similarity	Time(Min)
Same text	0.99	0.25
Different text	0.99	0.26

For the **BERT embedding**, the transformer library [121] is used, working only with the part that contains the tokenization of the text, and the attention mask is discarded since no predictions will be made. Since the texts do not have the same length, initially the distance of the longest text is determined and a padding is performed on the shortest text. The copy of the collapsed matrix to a single dimension is also made, so that it has the same shape and the Cosine similarity is calculated. The results are shown in Table 5.2.

TABLE 5.2: Cosine Similarity Calculation results with BERT Embedding.

Type of text	**Similarity**	**Time(Min)**
Same text	0.48	204
Different text	0.32	204

For the **GloVe embedding**, the gensim library [122] is used with a `glove wiki gigaword 50 model` which has the functionality of creating the similarity matrix. And the similarity of SoftCosine is calculated. The results are shown in Table 5.3.

TABLE 5.3: Soft Cosine Similarity Calculation results with GloVe Embedding.

Type of text	**Similarity**	**Time(Min)**
Same text	0.91	840
Different text	0.86	860

For Jaccard use no embedding is used. The results are shown in Table 5.4.

TABLE 5.4: Jaccard Similarity Calculation results.

Type of text	**Similarity**	**Time(Min)**
Same text	1.0	0.0006
Different text	0.99	0.0010

Among the best ones, the similarity calculated with simpler techniques such as Jaccard and Word2Vec can be highlighted. The difference between one technique and another is hardly noticeable, but in a matter of seconds of execution, it could be decided to use Jaccard.

5.2.4.3 Reuse phase

During this phase, it helps to increase the validity of similar cases retrieved. Firstly, the recovered cases are ordered from highest to lowest score. The best score is compared with the proposed threshold (50). If it exceeds it, it is shown to the user, otherwise a message is shown stating that the reliability of the recommendation is less than 50%.
For better compression, Figure 5.11 shows another case study where the threshold is not exceeded. For this, the following tweet is used: "Sonreír es lo más saludable que puedes hacer a diario".

```
Reuse phase:
Most similar is book with index:  68
Recommendation reliability: -50%

Answer:

- ElCazador
- LaDuquesaYElJoyero
```

FIGURE 5.11: In case the established threshold is not exceeded, an explanation message is displayed, although the two books with the best score are offered even if the threshold is not exceeded. Source: Own elaboration.

5.2.4.4 Revise phase

During this phase, the experts will determine whether or not the new case should be entered into the base case. For this, functions are implemented that allow validating the policies mentioned in the previous section. The operation of each of the implemented policies is shown below, where Figures 5.12 and 5.13 show the rejected cases.

```
Revise phase:
Enter validation (expert 1):n
Enter validation (expert 2):y
Enter validation (expert 3):n
There is not enough validation criteria to store this case in memory!!
```

FIGURE 5.12: First policy applied, if a single user is the one who thinks that it should be inserted, it will not be taken into account. Source: Own elaboration.

```
Revise phase:
Enter validation (expert 1):y
Enter validation (expert 2):y
Enter validation (expert 3):n
Expert 3 justify why you do not want to insert the case in the database
It's not the right book.
Thank you for your opinion, it has been stored.
```

FIGURE 5.13: Second policy applied, if a single user is against it, they must store the justification of why the case should not be saved. Source: Own elaboration.

The third policy is shown in Figure 5.14, where once the experts unanimously validate that the new case is apt to be saved in the base case, the next phase is continued.

```
Revise phase:
Enter validation (expert 1):y
Enter validation (expert 2):y
Enter validation (expert 3):y
The new case is suitable for storage!
```

FIGURE 5.14: Third policy applied, all experts agree that it should be inserted. Source: Own elaboration.

5.2.4.5 Retain phase

To better understand the data that is stored in our case memory, Figure 5.15 shows how the base case is found, with the initial data stored and in Figure 5.16 how it is found after storing the new case, once validated by the experts.

casesM > **libro**

_id	libro	texto	personalidad
6368cd96b557a4...	ElZapateritoDeG...	[16 elements]	ESTJ
6368cd96b557a4...	ElZapateritoDeG...	[33 elements]	INFP
6368cd96b557a4...	ElZapateritoDeG...	[14 elements]	ESTJ
6368cd96b557a4...	LaConciencia	[12 elements]	ESTJ
6368cd96b557a4...	LaConciencia	[5 elements]	ISFP

FIGURE 5.15: Base of cases with 150 samples, in this image only the last 5 are displayed. Source: Own elaboration.

casesM > **libro**

_id	libro	texto	personalidad
6368cd96b557a4...	ElZapateritoDeG...	[16 elements]	ESTJ
6368cd96b557a4...	ElZapateritoDeG...	[33 elements]	INFP
6368cd96b557a4...	ElZapateritoDeG...	[14 elements]	ESTJ
6368cd96b557a4...	LaConciencia	[12 elements]	ESTJ
6368cd96b557a4...	LaConciencia	[5 elements]	ISFP
637783e5a0618cf...	ElPerseguidor	[1 elements]	ISTP

FIGURE 5.16: base case updated with the new case inserted. Source: Own elaboration.

5.3 Non-fiction books

The entire process of creating the first part of the recommender based on Non-fiction books will be detailed below. Explaining the dataset to be used, the proposed tool, the experiments and the results obtained from them.

5.3.1 Dataset

The same Twitter dataset explained in the Fiction books section is used.

For the recommendation of Non-fiction books, another base case is filled in, solely with the data extracted from the GFK (4,807 samples). This database will contain the following characteristics: title, author, weekRelease, yearRelease, genres, format, price and units. Therefore, since the new case is based on the text introduced by a tweet to which a Non-fiction book is associated based on the theme it deals with, it is not feasible to store it as a case in the memory base. That is why, for these recommendations, they will not be evaluated or stored as new cases.

5.3.2 Proposed tool

In the case of recommendations for Non-fiction books, only two of the phases of a CBR will be taken into account, specifically, the Retrieve and Reuse phase, since the data stored in memory does not include the text itself, only characteristics of the book. This recommender starts with the problem entered, in this case, with the user's tweet which has previously been tagged to the topic that belongs to a Zero-Shot-Classifier. With this tag, the recommender is able to search for similar cases within memory, in the *Retrieve* phase. The recommended books are based on the cluster to which they belong for that subject. Finally, in the *Reuse* phase, the optimal case found is shown to the user. The general design of the system is shown in Figure 5.17.

5.3.3 Experimentation algorithms

5.3.3.1 Retrieve phase

For the recommendation of Non-fiction books, another perspective is used using NLP techniques, since these books already have a predefined theme to deal with. Therefore, this same theme should be associated with the users' tweets.

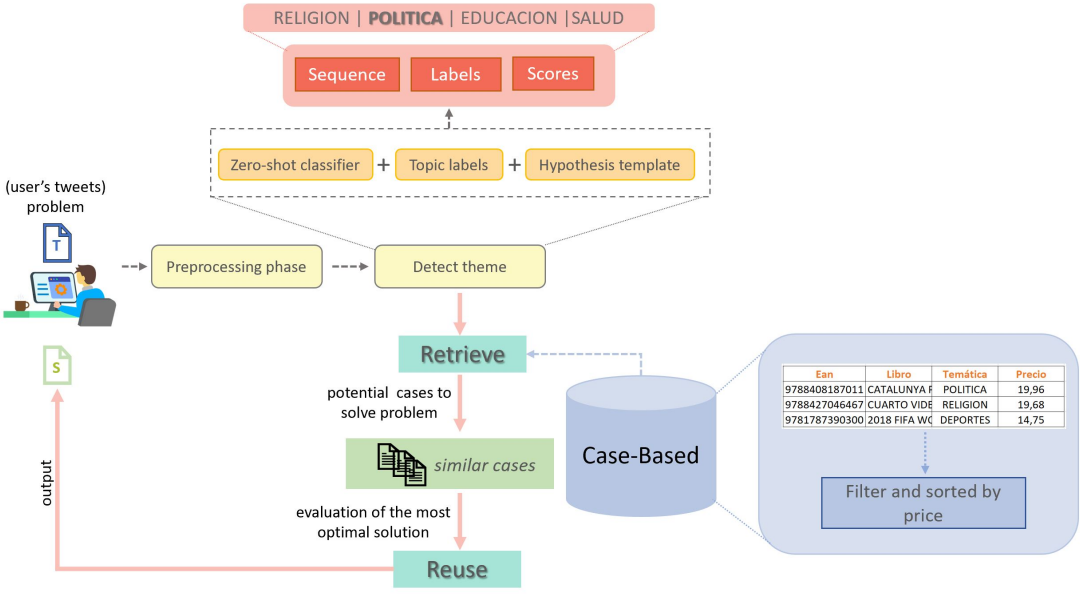

FIGURE 5.17: Case-based system architecture combined with NLP for Non-fiction books. With the Zero-shot-classifier task, the entered tweet can be tagged in one of the topics covered by nonfiction books. Once the theme to which it belongs is determined, the books of that theme are filtered within the base case and ordered by price, since price is the main factor when buying a book. Finally, the book with these characteristics is recommended. Source: Own elaboration.

Everyone knows that tagging tweets for certain topics can be a long and tedious process. And there is never enough data to train a model that covers all cases. Thanks to the advancement of the NLP field, zero activation learning can be counted on. This learning allows you to make a model perform a task for which it has not been explicitly trained, without relying on labeled data. With this model, the resource cost associated with tweet tagging can be reduced. The zero trigger method is based on natural language inference (NLI) [77]. It is the task of determining whether the given "hypothesis" follows logically from the "premise". In simple terms, it is to be understood if the hypothesis is true, while the premise is your only knowledge on the subject.

Until now, making a state of the art of the best models for Zero-shot-classification tasks, the following have been found:

- Facebook's Bart Large MNLI: This model is excellent for classifying texts in English, working with a potentially high number of candidate categories, while remaining fast and accurate. [123]
- Joe Davison's XLN Roberta Large XNLI: This model is perfect for non-English languages. The Cross-lingual Natural Language Inference (XNLI) corpus is the extension of the Multi-Genre NLI (MultiNLI) [79]. The model works in

more than 100 languages, and is particularly accurate in English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, Urdu. It has about the same latency as Bart Large MNLI.

5.3.3.2 Reuse phase

This phase performs an evaluation of the optimal solution and returns it as the proposed solution for the new problem. In this case, it is likely that the assigned category is the appropriate solution. To do this, the threshold is set at 50%. If it is passed, it indicates that the category classification carried out is reliable, displaying a message "Classification reliability: +50%". Otherwise, the category with the best score is assigned, but a message indicates that the recommendation is not significant enough, since the subject that interests the user cannot be determined.

5.3.3.3 Case study

With the case study presented below, it will be possible to examine the operation of the CBR with a real case. In Figures 5.18 and 5.19 the internal flow of each of the phases of the proposed system.

```
                    BRAIN L - Expert system
New case: En el directo del pasado lunes estuve hablando de como deberiamos usar las redes sociales en el periodismo deportiv

Retrieve phase:
sequence:  En el directo del pasado lunes estuve hablando de como deberiamos usar las redes sociales en el periodismo deportiv
labels:  ['DEPORTES', 'VIDEOJUEGOS', 'EMPLEO', 'EDUCACION', 'ECONOMIA GENERAL', 'CINE', 'POLITICA', 'SALUD Y BIENESTAR', 'MUSICA', 'RELIGION', 'FOTOGRAFIA', 'FINANZAS']
scores:  [0.9704673290252686, 0.006882543209940195, 0.0040464154444634914, 0.002811947138980031, 0.0024431340862065554, 0.0023016559425741434, 0.0020634892862290144, 0.00
```

FIGURE 5.18: Recovery phase. The ranking score for each of the topics is displayed. Source: Own elaboration.

The previous image shows how a new case is introduced, specifically, a tweet. When searching for similar cases internally, the system will return the most similar books and the similarity score.

```
                    BRAIN L - Expert system
New case: En el directo del pasado lunes estuve hablando de como deberiamos usar las redes sociales en el periodismo deportiv

Reuse phase:
Recommendation reliability: +50%

Answer: The book RECUERDOS GRANAS  is recommended. Belonging to the theme: DEPORTES
```

FIGURE 5.19: Reuse phase. Obtaining the maximum score. It is analyzed if it exceeds 50% confidence that the recommendation is correct. If you get over it, that's the case. The reader is shown the recommended book, including the subject in which it has been classified according to its writing on Twitter. Source: Own elaboration.

5.3.4 Results

Once the operation of the system with the case study has been visualized, the results of the experiments carried out are presented.

5.3.4.1 Dataset

Non-fiction books

In the case of Non-fiction books, before storing them in our case database, a small data processing is carried out. The total number of units sold is calculated and the sales channel and region characteristics are eliminated; since it did not provide significant information for the recommender.

5.3.4.2 Retrieve phase

During the model, it takes `xlm-roberta-large` and fits it to a mix of NLI data in 15 languages. Once the pretrained model is downloaded, it must be saved and loaded along with a tokenizer that we will create with the pretrained model itself. With the purpose, that during the call to the transformer's pipeline, the tweet can be tokenized with the same inscription in which the training data of the pretrained model was tokenized.

Once the classifier has been created, the candidate labels are introduced, in this case the same themes stored in the base case. Obtaining as a result the labels again and the score of each one of them, leaving the system with the best of all. For a better understanding, in Figure 5.20 can observe an example of labeling a tweet.

```
{'sequence: 'En el directo del pasado lunes estuve hablando de como
deberiamos usar las redes sociales en el periodismo deportiv',
'labels': ['DEPORTES', 'VIDEOJUEGOS', 'EMPLEO', 'EDUCACION', 'ECONOMIA
GENERAL', 'CINE', 'POLITICA', 'SALUD Y BIENESTAR', 'MUSICA',
'RELIGION', 'FOTOGRAFIA', 'FINANZAS'], 'scores': [0.9704673290252686,
0.006882543209940195, 0.0040464154444634914, 0.002811947138980031,
0.0024431340862065554, 0.0023016559425741434, 0.0020634892862290144,
0.002054498065263033, 0.0019259806722402573, 0.0018027224577963352,
0.0016323616728186607, 0.0015679446514695883]
```

FIGURE 5.20: Zero-shot-classification enforcement output in a tweet. Source: Own elaboration.

The previous image shows the output of the classifier which has three components: the sequence that shows us the entered text; the set of candidate labels, in this case thirteen labels corresponding to the same themes of the Non-fiction books;

and finally a set of thirteen corresponding scores between 0 and 1 that represent a probability distribution of the probability of linking to said theme.

In this particular example, the result shows that the tweet belongs to the 'DEPORTES' (*SPORT in english*) theme, as it reaches the highest confidence score of belonging to this category with 0.97.

Of the 7,535 tweets, around 5,018 scored above 50% for a particular tag and are therefore ranked. The rest of the scores obtained that were too low in all the labels and would probably have been classified as "Other", which means that the current classification of Non-fiction book themes could be further disaggregated into other sub-themes, since it is not being completely inclusive. As can be seen in Figure 5.21, the most popular topic is "DEPORTES" (SPORTS in English), followed by "EMPLEO" (JOB in English). The least popular topic is "ECONOMIA" (ECONOMY in English), followed by "FOTOGRAFIA" (PHOTOGRAPHY in English).

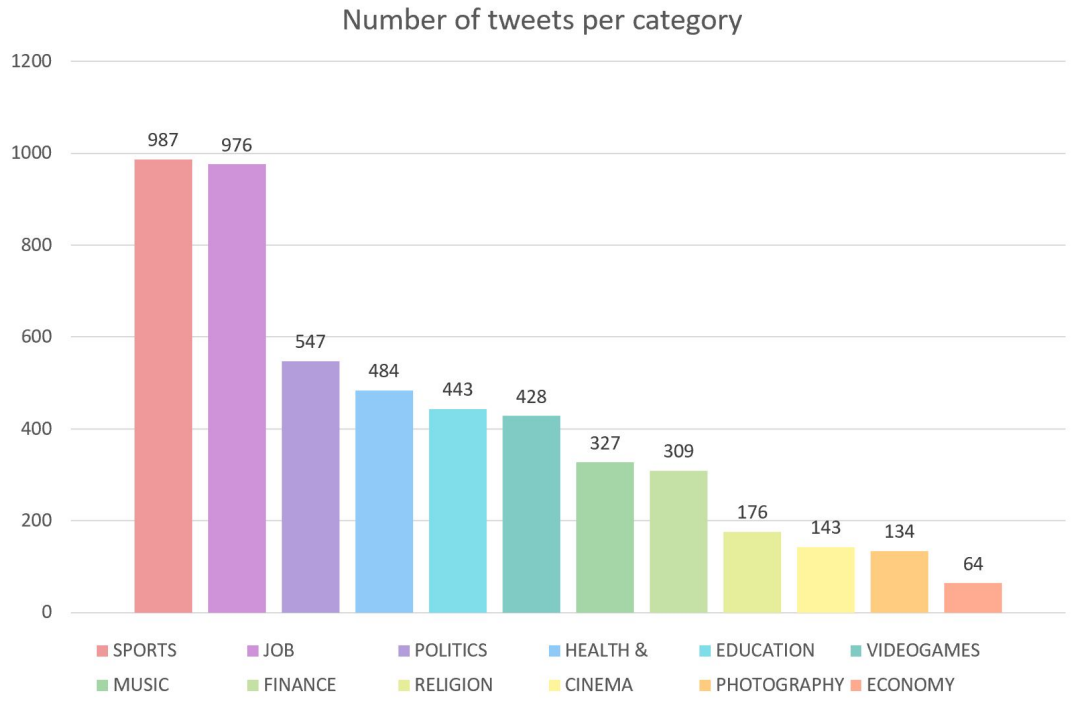

FIGURE 5.21: Number of tweets per category correctly classified with zero-shot classifier. Source: Own elaboration.

5.3.4.3 Reuse phase

During this phase, it helps to increase the validity of similar cases retrieved. Firstly, the recovered cases are ordered from highest to lowest score. The best score is compared with the proposed threshold (50). In the case of not exceeding the threshold, a message is displayed that a recommendation is made but it does not have a significant reliability, since it has not been possible to correctly determine the topic of interest. It is important to mention that once the theme to which the tweet belongs is obtained, all those books belonging to this theme are filtered within the base case and are ordered by price, since it is the characteristic most correlated with the sales units of the books. For better understanding, Figure 5.22 shows another case study where the threshold is not exceeded.

```
                    BRAIN L - Expert system
New case: La industria y las autoridades se reunen en Tarragona para impulsar el programa Operation Clean Sweep contra las f

Reuse phase:
It has not been possible to determine the topic that interests you. However, we recommend two types of books that may interest you.

Recommended book:  NORMAS APLICABLES A LA ESTETICA
Belonging to the theme: SALUD Y BIENESTAR

Recommended book:  CUADERNO DEL ALUMNO. GESTION AUXILIAR DE LA CORRESPONDENCIA Y PAQUETERIA EN LA EMPRESA (CERTIFICADOS CERTIFICADOS DE PROFESIONALIDAD). OPERACIONES
Belonging to the theme: EMPLEO
```

FIGURE 5.22: In case the established threshold is not exceeded, an explanation message is displayed, although the two books with the best score are offered even if the threshold is not exceeded. Source: Own elaboration.

5.4 Summary

This section contributes to personalized book recommendation. Focused on the case of fiction books on the text itself, rather than on the sales history. This will allow the reader to feel more empathetic with the book, since the characters found in it reflect psychological characteristics similar to those presented by the reader, given their way of writing. And in the case of Non-fiction, it will be based on the topics that the book deals with. With this, reading is also encouraged, since the reader will find similarities with the characters of Fiction works or will be interested in the topic that Non-fiction books deal with.

To put this created system into production, it is necessary to highlight that the recommendation of a book to a user would be applied 120 hours after the information from their tweets was collected. The average level of tweets per day is 4 or more, if they are prolific users. For this recommendation, a list of books ordered by the values of the text similarity distance will be generated.

Part III

General conclusions

Chapter 6

Discussion

"In any discussion, it is not a thesis that is defended but oneself."
-Paul Valery-

Throughout this chapter we will highlight our contribution to editorial world. The limitations encountered during the development of this project will be discussed. As well as, on the ethical and social considerations that must be taken into account in Artificial Intelligence solutions.

6.1 Introduction

This chapter presents, summarizes, interprets and extrapolates the results obtained in each of the previous chapters related to the publications [6][81][107][117], which have been presented in this document. Their implications and limitations are analyzed, and they are confronted with the proposed hypothesis, considering in the corresponding cases the perspective of other authors who present works related to the subject in question.

6.2 Discussion

Next, the results obtained in each of the publications will be discussed.

6.2.1 Readers permanence

The system proposed in Chapter 1 and in the study by Martin et al. [6] served as a previous basis for the Doctoral Thesis. This system allows segmenting the news into three types of clusters according to the length of stay of the readers: low, medium and high permanence. Demonstrating that in 88% of cases, it is possible to predict how much time users spend reading news. This allows us to understand the characteristics that a reading must have (in this case, digital news) to be of interest to the reader. Although it is true that digital news does not have the same characteristics as books, the importance of the text itself is highlighted. What motivates us to investigate more in the world of Natural Language Processing. However, this study allowed us to open new hypotheses based on the results obtained: a) The quality of the audio in digital news influences the permanence of the reader; b) What characteristics can be integrated into the medium permanence cluster to improve this group and convert it into the high permanence cluster and c) Trending topics influence a reader's permanence in a specific article. The first two being very specific to the digital press domain, but the last one encourages us to investigate the topics that interest the reader and the readings we offer. Therefore, this analysis is a contribution to the continuous improvement of a piece of news before its publication based on the dimension of the content. It also encourages new authors to look for new indicators that optimize the communication between the journalist and the reader.

6.2.2 Sales predictions

With the hybrid model proposed in Chapter 3 and in the study by Martin et al. [81], it is possible to predict the number of copies of books before publication. However, we are aware that in order to reach the final validation of the improvement in the number of copies to print predictions through the proposed segmentation, it is necessary to take into account multiple factors, in which we will stop to analyze them. In the first part, it is a matter of deepening the analysis of the marketing actions that will be carried out in each of the sales to increase their impact. Obviously, these actions are closely linked to the social network factor. They must be used to empathize with the reader, create and care for the link with him. Of course, the risk of the topics that deal with the reader or the way in which they do it, will be directly related to the author and not to the marketing department. One way to take care of the link by the marketing department is to offer promotions on a case-by-case basis. Another factor to take into account to verify the operation of the model is to extend the period of analysis, observe patterns of behavior over the years and the situation that exists at that time. Closely linked to the period of analysis factor is the type of social networks that exist or are in fashion, since technological advances allow us to insert new networks within society. During the analysis of results for the preparation of this hybrid model, it is important to emphasize the phenomenon of interpretations from the business point of view, since until the date of creation of this model the limits established by the experts used to be quite arbitrary. Thus, it was possible to observe, with the results of the natural form of the data, following a totally different distribution and in which it is not so easy to distinguish them, hence many are grouped into a single cluster. Probably because there are other variables that the experts could not share for various reasons, from not being an easy variable to extract (because of the complexity involved) to confidentiality issues.

6.2.3 Character profile

With the system proposed in Chapter 4 and in the Martin et al. [107] study, more promising results can be obtained for determining a character's psychological profile. The initial detection of the characters was carried out taking into account the lessons obtained from similar works [52]. From the graphics (in the final visualization) it was possible to detect not only the real number of characters in the work, but also that as a whole they favored the discovery of the type (choral or not) of literary work. Finding that until now this was not digitized, and that it serves as a quality or key indicator for the publishing sector, especially

for publishers in their determination of whether the work is publishable or not. The approach of discovering the belonging of the personal name to one of the implicit characters in a range of 300 words, where a relationship between the characters is established as guaranteed by Elson [57], allowed us to have 80% more functionalities than the characters themselves (see Table 4.3) and thus improve the creation of their profile. Our results are high, probably because only he and she were used, although we leave the inclusion of forms of omission of these pronouns as a problem for future work. For example, in the following sentence where the personal pronoun is omitted, since in Spanish the pronoun is usually implicit in the verb:

"Anderson se presentó a un tercer mandato en las elecciones generales de Alberta de 1986. Ganó esa elección derrotando cómodamente a otros tres candidatos, pero su porcentaje de votos disminuyó sustancialmente. Después de las elecciones, el primer ministro Don Getty nombró a Anderson como Ministro de Cultura. Él ocupó esa cartera hasta 1987, cuando pasó a la cartera de Asuntos Municipales.". In English it means: *"Anderson ran for a third term in the 1986 Alberta general election. He won that election handily defeating three other candidates, but his share of the vote was substantially reduced. After the election, Prime Minister Don Getty appointed Anderson Minister of Culture. He held that portfolio until 1987, when he moved to the Municipal Affairs portfolio."*

We agree with the approach of the studies by [58] where they reach 50% accuracy in determining personalities in text, although no more information about the character itself has been incorporated. Therefore, including this information in our research allowed us to be 8% more successful than their results. This is significantly relevant, given the limitations that NLP has in Spanish texts. Although there are many advances in NLP, we find a long way to go in the Spanish language. In particular, in creating datasets that are robust enough for model learning. But above all, carry out more accurate studies of personality, and for this reason it is required: a) A larger set of data and a Spanish suitable for the personality labels proposed by Myers-Briggs, b) A tool that allows the labeling of the Named Entity Recognition in a more automated way, c) Increase the study of personality. until the determination of whether the character is the hero or the villain of the work, d) Establish and identify the relationships between the characters and e) Extend the analysis of the literary work to the rest of the keys proposed by Vladimir Propp [105].

6.2.4 NLP-based recommender system

With the system proposed in Chapter 5 and in the study Martin et al. [117], Fiction books begin to be recommended more based on what the reader transmits, than with the sales history. For this we have based ourselves on the idea of Wu et al. [73] study, since of all the related works reviewed to date it is the one that most resembles the idea that we wanted to work on in this research. However, his idea was based more on ontologies than on the text itself, and it is the main contribution we make with our work. Based on the semantics of the texts, the objective is to recommend the book that best suits the reader, and thereby also determine the psychological profile of the reader, thanks to previous work done by Martin et al. [107].

During experiments it has been shown that although the most advanced embedding technique is BERT, it is not yet ready to compute text similarity. The resulting data is unreliable and has a high runtime cost. It is shown that with initial techniques in NLP such as Jaccard or Word2Vec good results can be obtained, such as 83% in texts whose words are different, but mean the same thing; with an almost instant execution time. Obviously, we are aware that there is still a long way to go, such as: a) expanding the memory of cases, b) creating a focus group that allows us to use their tweets and then offer us their opinion if the recommendation was interesting or not, and finally, c) have the vital experience of the experts to validate the new cases that enter our system. However, with this work we want to sow a seed of how to improve recommendation systems in the literary sector, where currently it is only recommended based on historical sales.

Finally, to close the book recommendation cycle, the creation of another database is proposed. On this occasion only with the Fiction books, which have the theme as a characteristic. This allows us to recommend books based on the interests of the public that has been present since the beginning of the investigation of this thesis project. The results on this occasion may continue to be better, given that the topics of interest can be expanded and many of the existing ones can even be disaggregated. With this, we would guarantee that the recommendations provided are adjusted much more.

6.3 Limitations

Although it is considered by the author that with the results of the thesis work it has been possible to answer the proposed hypothesis, different limitations have been found during its development, which are set out below.

1. Volume and data collection
2. Computational power
3. There is no free text annotator tool for Name Entity Recognition and Part Of Speech, which is complex for creating your own Spanish language datasets.

6.4 Ethical and social considerations

As the last part of our discussion (and no less important), it is necessary to highlight that no personal data has been processed for the analysis of the reader's interests, thus complying with the General Data Protection Regulation (RGPD). While the work done makes it possible to personalize sales, it also validates the fact that social media needs to be handled right. Despite the fact that the recommend system created is an advisory tool for the personalized tastes of potential customers (based on their own tweets, instead of purchase history); the excessive use of these platforms would leave humans without the main tool of living beings, reasoning. This leads us to ask ourselves different questions: Where are the limits of Artificial Intelligence? How can we include ethics in the behavior of a model?

Finally, with this chapter we highlight the benefits of the system, a critical analysis of the results is carried out, the limitations for its development are exposed and continuous improvement is considered by applying our ethical sense in this type of work.

Chapter 7

Conclusions

"Life is the art of drawing sufficient conclusions from insufficient data."
-Samuel Butler-

This chapter presents a high-level vision of the design that the proposed tool will have. The details of each of the designed parts and the contributions with each one of them to the scientific community are exposed and the results obtained from the enhanced experimentations are analyzed.

7.1 Introduction

During this chapter, the conclusions are exposed after finishing all the experimentation carried out to answer the research question. Future work and contributions made to the scientific community are exposed. The calendar followed throughout these 3 years is shown. And it emphasizes the lessons learned both professionally and personally.

7.2 Conclusions

In this thesis, the following hypothesis was proposed as a starting point: *"It is possible to build a book recommendation system that adjusts to the needs of the reader and allows him to enrich his reading and satisfaction."*, which led to the following general objective: *Design a book recommender based on the reader's behavior that encourages greater reading and satisfaction.*
This main objective has been fulfilled as reflected in Chapters 3, 4 and 5 which show the different analyzes and experimentation that have been carried out into account for the proposal and construction of the tool. The results in each of the sections are validated, seeking to adjust them as much as possible to the needs of the readers.
To achieve the general objective, a set of specific objectives were proposed, which have been fulfilled during the development of this doctoral thesis. The scope of each of them is detailed below:

1. Segmenting the recommendation according to the literary genre: This first objective is of vital importance for the development of the thesis, since the sales and/or the impact of a Fiction book are not analyzed with the same characteristics as that of Non-fiction. For this reason, an analysis is first carried out on the part of the Non-fiction books, focusing on sales predictions; and in a second part in the Fiction books, focusing exclusively on the analysis of the texts.

2. Influence of an author in social networks: During Chapter 3 an analysis is carried out and the reasons why it is vitally important to include social networks as a variable to analyze in the sales prediction are detailed. Basically because sales have a high correlation with the public, and currently society tends to focus its relationships based on social networks. The results have shown that the prediction of the number of copies to print improves significantly

if automatic segmentation methods are used, reaching between 94-100% accuracy in each one of the segmentations, respectively. Being the segmentations: a) class 1, corresponding to the lowest sales; b) class 2, intermediate low sales; c) class 3, intermediate high sales and finally d) class 4 super high sales (known in the publishing sector as "Bestseller"). Another implicit contribution of this type of analysis is that once they are validated with a much larger amount of data, more sustainable consumption and production patterns can be guaranteed within the framework of the action plan to implement the 2030 Agenda [124].

3. Create a character profile of a work: During Chapter 4 a model is built that allows the creation of a psychological profile of characters of a literary work, obtaining quite promising results, with a Minimum 8% improvement in determining a person's personality. In the related work by Argamon [58] they are able to predict the personality of the authors by 50%, and with the proposed model they can predict the personality of the characters by 54%. It is important to note that this accuracy rate for determining personality type from a single short text is quite significant, as personality is difficult to demonstrate without specific questions. With these results, readers feel more identified with the personality traits of the characters in the work they are reading because they are very similar to those of the readers.

4. Design a model capable of not losing any information about the character: This objective is closely linked to the previous one, since the work presented in Chapter 4 prevents the loss of relevant information that they may have sentences that use personal pronouns, since it is 80% guaranteed to find the character that the personal pronoun refers to. Finally, to close the analysis of the characters of the work of Fiction, it contributes during this experimentation to the definition of whether a work is choral or not, an interesting factor for determining whether a work should be published or not. In summary, both the psychological profile of the character and the determination of whether the work is choral or not, will allow the publishing sector to create indicators for the digitization of literary analysis that is currently carried out manually by experts.

5. Develop a recommendation platform: This objective is closely linked to the first, since recommendations are made based on gender. Finally, a Case-Based Reasoning system is developed using Natural Language Processing techniques. Achieving promising results, despite the limitations that exist. Being the main contribution of a book recommendation system focused on the text itself and not on historical sales, given that to date no complete CBR system has been found that focuses on this type of recommendation focused on the publishing environment.

7.3 Futher work

Although the objective of this thesis has been fulfilled, there is still much work to be done, especially in the field of Natural Language Processing in Spanish. Although a first approach to the use and implementation of case-based recommendation systems has been made, this model can be refined and extended to support additional features. In addition, it leaves open different lines of research to complete and improve the defined methods and tools, such as:

- Put this study into production once it has been validated with sufficient data. To put this system into production it is required: a) Collect the tweets from the users, transform them to the input data format to the system (grouping of tweets, preprocessing), b) Execute the model and finally, c) Generate a report of the list of recommended books, so that the Marketing department can use it to carry out online campaigns in the banners of Twitter itself.
- Expand this type of analysis to other longer works of Fiction, such as novels.
- Add new external variables such as: reading time of a focus group; analyze reader reviews of purchased books, among many others.
- Expand the recommender and take into account authors, genres, among others.
- Increase validation policies.
- Create an interface that is friendlier to the end user.
- Validate the capabilities of the CBR tool through MLOps [125], which is a set of practices that aim to implement and maintain machine learning models in production reliably and efficiently. Allowing to expand the use of the tool in other retail fields, such as textiles, music and movies, etc.

7.4 International Contribution and talent pool

This section will detail our contributions to journals, congresses and publications of the scientific community during the development of this project.

1. At the Intelligence Conference Systems we have contributed with a new hybrid decision support tool. The one that allows to predict the impact of digital news. In this work, we have validated the proposed tool:

   Martín Sujo, J. C., Golobardes i Ribé, E., Vilasís Cardona, X., Jiménez Ruano, V., & Villasmil López, J. (2022). Correction to: SmartData: An Intelligent Decision Support System to Predict the Readers Permanence in News. In Proceedings of SAI Intelligent Systems Conference (pp. C1-C1). Springer, Cham.

   Abstract: This article proposes a hybrid intelligent system based on the application and combination of Artificial Intelligence methods as a decision support tool. The objective of this study is to exploit the advantages of the constituent algorithms, to predict the permanence rates of readers in news from a digital media. With this, the editor will be able to decide whether to publish a news item or not. To evaluate the effectiveness of the hybrid intelligent system, data from a reference digital media is used. In addition, a series of performance metrics is calculated, where 88% effective is demonstrated with the predicted results.

2. In the Women Special Issue, we have contributed a new book sales prediction tool. In addition to the contribution itself, we highlight the participation and contributions made by women in the scientific world, to the point that this work has been used as a reference in the article [126] and as well as, it is part of a chapter of the book created by the same authors [127]. In this work, we have validated the proposed tool:

   Martín Sujo, J. C., Golobardes i Ribé, E., & Vilasís Cardona, X. (2021). CAIT: A Predictive Tool for Supporting the Book Market Operation Using Social Networks. Applied Sciences, 12(1), 366.

   Abstract: A new predictive support tool for the publishing industry is presented in this note. It consists of a combined model of Artificial Intelligence techniques (CAIT) that seeks the optimal prediction of the number of book copies, finding out which is the best segmentation of the book market, using data from the networks social and the web. Predicted sales appear to be

more accurate, applying machine learning techniques such as clustering (in this specific case, KMeans) rather than using current publishing industry expert's segmentation. This identification has important implications for the publishing sector since the forecast will adjust more to the behavior of the stakeholders than to the skills or knowledge acquired by the experts, which is a certain way that may not be sufficient and/or variable throughout the period.

3. The world of NLP has been revolutionized in recent years, but there is still a long way to go in different languages, especially in Spanish. In the prestigious Elsevier magazine we contribute with a new psychological profile creation tool for the characters of works of Fiction in Spanish. In this work we have validated the proposed tool:

   Martín Sujo, J. C., & Golobardes i Ribé, E. Personality Profile of Fictional Characters in Books Using Natural Language Processing. Submitted to Computer Speech & Language - Elsevier

   Abstract: This study focuses on the creation of a character profile of a fictional literary work in Spanish with Artificial Intelligence techniques. A tool is presented and designed that avoids the loss of information caused by the reduction of cacophony in texts. The Bidirectional Transformer Encoder (BERT) layer is incorporated as embedded in real tasks with NLP applications, thus guaranteeing that the model is capable of understanding the general context of the text and not just a part of it. The available tool is evaluated, showing significantly better results compared to pretrained models in character detection and association of the personal pronoun between two characters. It contributes to integrating multidisciplinary professions such as psychology, literature or artificial intelligence, showing promising results, even achieving a minimum improvement of 4% in predicting a character's personality compared to literature.

4. As a last contribution during the development of this project, a last tool is proposed for the Elsevier magazine that allows the recommendation of a book based on the text itself. A considerable contribution in the publishing sector, since the authors has not found a similar work on this subject to date. In this work, we have validated the proposed tool:

   Martín Sujo, J. C., & Golobardes i Ribé, E. BRAIN L: A book recommender system. arXiv preprint arXiv:2302.00653. In preparation

Abstract: Book sales in Spain have fallen progressively, which requires urgent changes to optimize the sales process as much as possible. This research proposes a new expert system Base of Reasoning in Artificial Intelligence with Natural Language (BRAIN L) focused exclusively on the publishing industry. Integrate the new field of knowledge of AI, NLP with CBR techniques for book recommendation. A model was developed to retrieve similar cases/books supported by NLP techniques for decision-making. In addition, policies were implemented to keep the model evaluated by expert reviews, where the system not only learns with new cases, but also with cases where the value of truth is implicit.

7.5 Schedule/Work plan

In this section, you observe a series of tasks and time management that summarize this thesis work. Like any project, this one needs a work plan and specific planning. Table 7.1 shows the development of the tasks of the doctorate. Likewise, the details of each of the tasks carried out can be seen below.

TABLE 7.1: PhD task devolopment

Month	0	3	6	9	12	15	18	21	24	27	30	33	36
T1													
T2													
T3													
T4													
T5													
T6													

1. Task 1 (T1): During the first 3 months, the global development of this thesis was planned, based on the prePhD work carried out with the permanence of readers in news. Therefore, we began with the capture of requirements to start the first of the experiments. An exploratory analysis of the data was performed.

2. Task 2 (T2): During the following 9 months, the execution of the first experiments began, searching for the optimal algorithms (including the combination of several) for the prediction of sales of Non-fiction and Children's/Youth books. During this period, the participation in the 1st international conference "Intelligence Conference System" takes place with the work of prePhD, which is published within a collection of academic articles.

3. Task 3 (T3): During the following 6 months, an analysis and study of the aspects that are taken into account for the publication of a Fiction book during the editorial process is carried out. Focusing the work on one of 12 aspects: The characters of the literary work.Also begin, during this period, the solid study of the world of Natural Language Processing. A state of the art of the NLP is made in Spanish. During this period, the second publication is produced with the work carried out during the first experimentation of this thesis.

4. Task 4 (T4): During the following 6 months, the execution of the second experimentation carried out during this doctoral process is carried out. Thus contributing to the scientific community to the integration of dissimilar branches such as: Artificial Intelligence, Literature and Psychology. The results allow a third publication in the prestigious Elsevier magazine.

5. Task 5 (T5): During the following 3 months and according to the global planning, a state of the art of the current recommender systems is carried out. Conceptd such as Case Based Reasoning are studied during this stage.

6. Task 6 (T6): During the following 6 months, the results of the previous experiments serve as the basis for the memory cases of the recommender system that is developed in this period. This allows a publication to be made with the contributions made with this expert system. And preparation of the final thesis.

7.6 Lesson learned (Hard skills)

Before concluding this chapter, it is important to comment on the skills acquired during this doctoral thesis process:

- Perform statistical analysis of the results and refine the models.
- Design and develop natural language processing systems.
- Define appropriate datasets for language learning systems.
- Implement algorithms and tools suitable for NLP tasks in Spanish.
- Gain BERT transfer learning experience.
- Improve oral expression in presentations.
- Write articles.
- Scientific rigor.

7.7 Consolidate skills (Soft skills)

Natural Language Processing has begun to pay attention to the great diversity of languages spoken in the world. And with this a window of possibilities opens, however, it also implies important technological and social challenges. In this section, he lists the soft skills I did to help me develop this complex topic during this 3-year journey:

- Self-motivation
- Overcome obstacles
- Adaptation

Appendix A

Sales Predictions Experiments

A.0.1 Silhouette index evaluation metric

Another evaluation metric Silhouette index is applied, with this coefficient the ideal number of clusters is indicated. As shown in Figure A.1, the numerical results indicate that the optimal number would be 4, since it presents the highest value. Furthermore, if the Figure A.2 is analyzed in said group , it is observed that the object is well matched with its own cluster.

```
For n_clusters = 2 The average silhouette_score is : 0.6797831674097246
For n_clusters = 3 The average silhouette_score is : 0.7031445769244088
For n_clusters = 4 The average silhouette_score is : 0.7186692590088685
For n_clusters = 5 The average silhouette_score is : 0.7062480357527594
For n_clusters = 6 The average silhouette_score is : 0.6245154711697105
For n_clusters = 7 The average silhouette_score is : 0.6288735555944069
For n_clusters = 8 The average silhouette_score is : 0.5792759420916422
For n_clusters = 9 The average silhouette_score is : 0.4633189929975322
For n_clusters = 10 The average silhouette_score is : 0.4486092585281737
```

FIGURE A.1: Silhouette score. Source: Own elaboration.

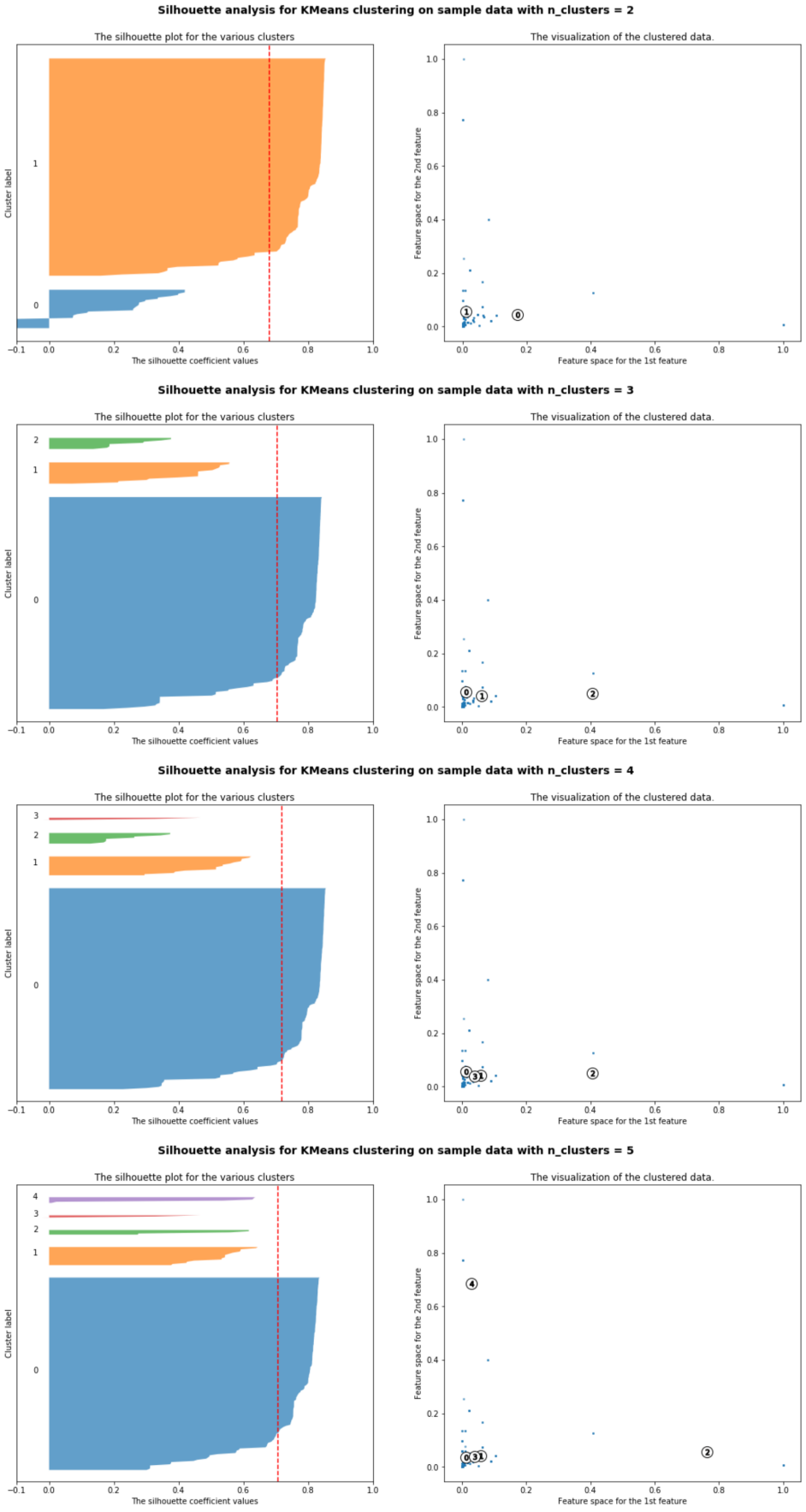

FIGURE A.2: Silhouette metric for the first 5 groups. Source: Own elaboration.

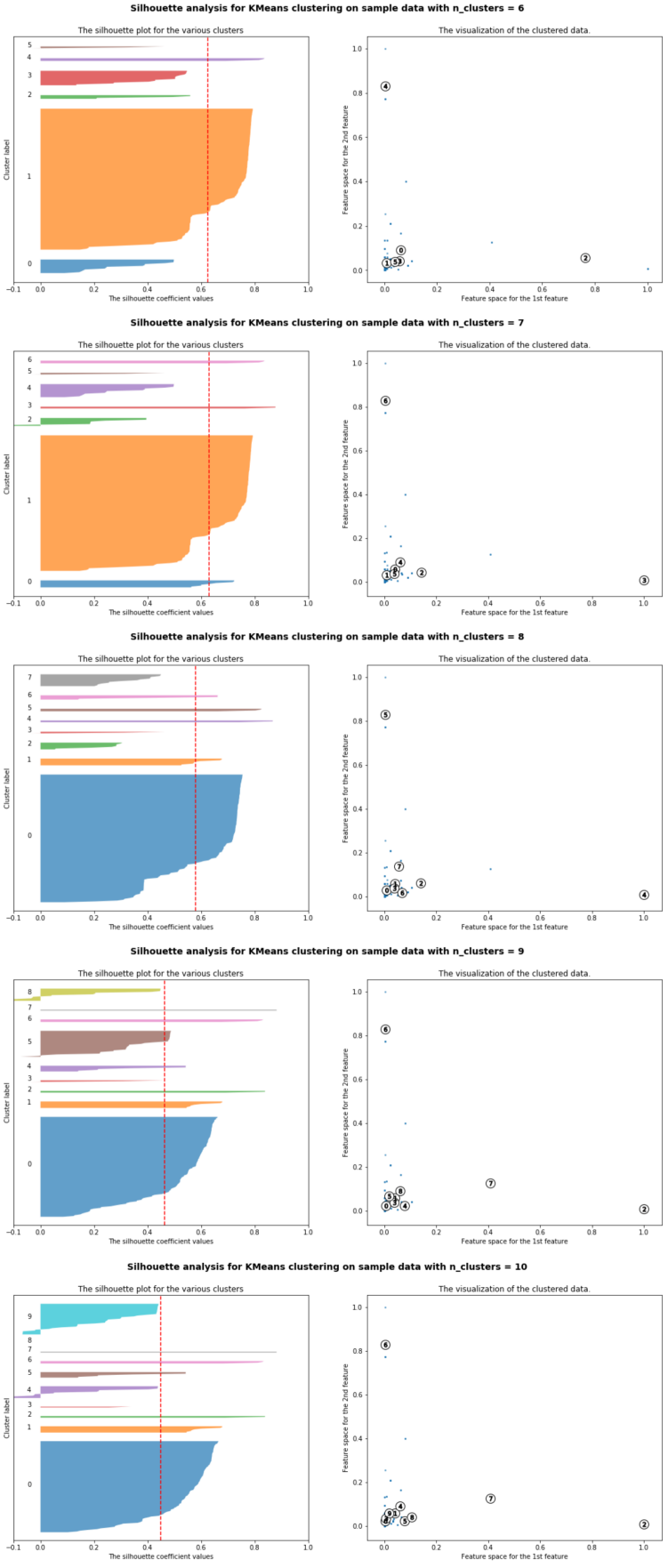

FIGURE A.3: Silhouette metric for the following 5 groups. Source: Own elaboration.

A.0.2 Expert segmentation classes analysis

In Figure A.4 the value 9.37 is the prediction of the model for Class 1 according to the characteristics that were input to it. The parameters that are in red are the ones that make the prediction have a higher value, while the parameters that are in blue make the prediction have a lower value. It can be clearly seen that the variable `value` has high relevance, which causes the prediction to increase in value from the base value; followed by `numberMentions`, `weekRelease`. The impact of these variables in this category are analyzed in more detail below.

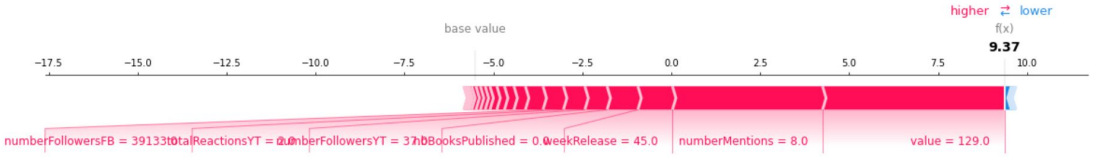

FIGURE A.4: Force Plot to visualize the importance of each input parameter in the model prediction for Class1-Low sales. Source: Own elaboration.

In Figure A.5, belonging to Class 1 - Low sales, it can be seen that the features are ordered by Tree Shape and, therefore, it is observed that the feature `value` it is mostly associated with low values and positive with respect to sales. Implying that people buy more books if the price is low. On the other hand, the characteristic `numberPublicationFB`, which is the one with the least relevance, indicates that it has more low and negative values with respect to sales. Implying that people buy fewer books, when there are fewer publications on the Facebook social network, which can serve as an indicator for the commercial department to promote more books in this category on this social network.

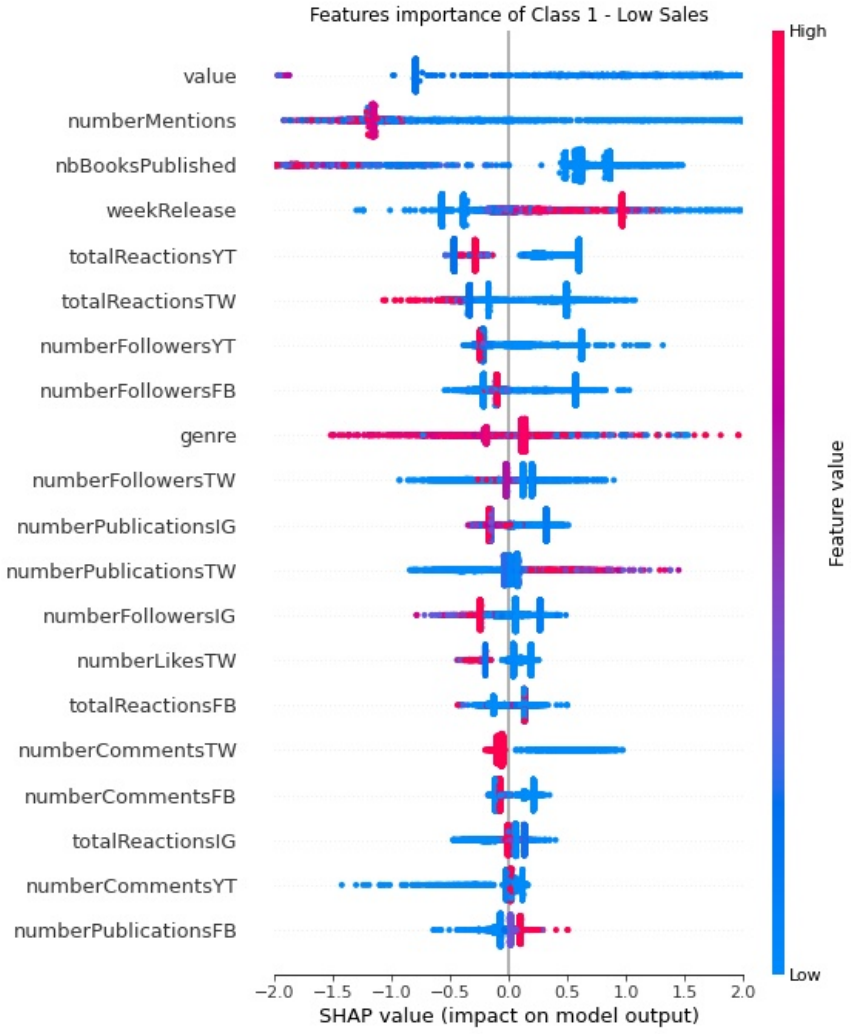

FIGURE A.5: Summary Plot to see the relationship of all the variables with the model for Class1, and the impact they have. Source: Own elaboration.

If Class 2 - Intermediate Low Sales is analyzed, it can be seen in Figure A.6 an effect totally opposite to the previous class. The variable `value` has high relevance, but in this category the prediction decrease in value from the base value; followed by `numberFollowersYT`, `numberFollowersFB`. The impact of these variables in this category are analyzed in more detail below.

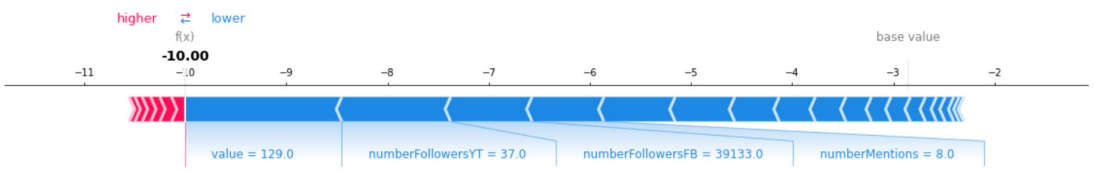

FIGURE A.6: Force Plot to visualize the importance of each input parameter in the model prediction for Class 2. Source: Own elaboration.

In Figure A.7, belonging to Class 2 - Intermediate low sales, it can be seen that the characteristic `value` it is mostly associated with low and positive values with respect to sales. Which implies that people buy more books if the price is low. Although it can also be seen with these intermediate sales, the `value` presents high and negative values with respect to sales, indicating that people buy fewer books if the price is more expensive. On the other hand, the feature `numberPublicationFB` tells us that people buy more books, with fewer publications on this social network. If the characteristics of the other social networks are observed, it is observed that people buy more if they have, for example, more publications on Instagram or Twitter. Therefore, it is suggested to the commercial department to take advantage of this type of network for their campaigns.

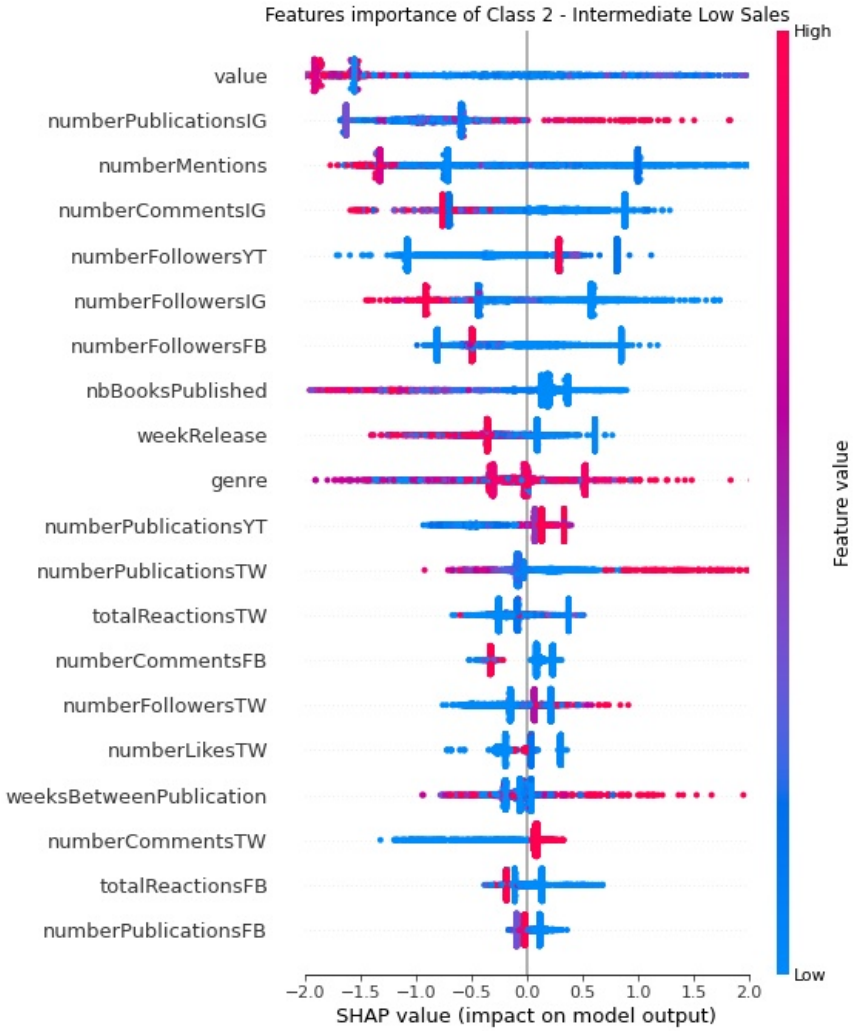

FIGURE A.7: Summary Plot to see the relationship of all the variables with the model for Class 2, and the impact they have. Source: Own elaboration.

In Figure A.8 the feature `numberPublicationsFB` has high relevance, but in this category the prediction decrease in value from the base value; followed by `value`, `numberMentions`. The impact of these variables in this category are analyzed in more detail below.

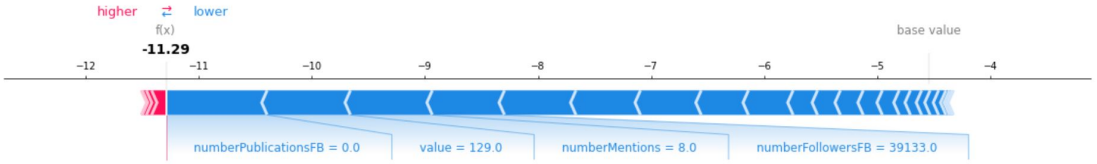

FIGURE A.8: Force Plot to visualize the importance of each input parameter in the model prediction for Class 3. Source: Own elaboration.

In Figure A.9, belonging to Class 3 - High intermediate sales, it can be seen that in the characteristic `value`, there is a variety of book purchases, on the one hand there is a majority (pink color) that buys fewer books, if the prices are very high. Although if the distribution is analyzed you can also find atypical cases, which do not care about this factor, and buy more books despite having a high price. If the variable `weekRelease`, the one with the least relevance for this category, is analyzed, outliers are again found, but a large part (pink color) indicates that people will buy more books, if there is a greater distance from the release dates of the same.

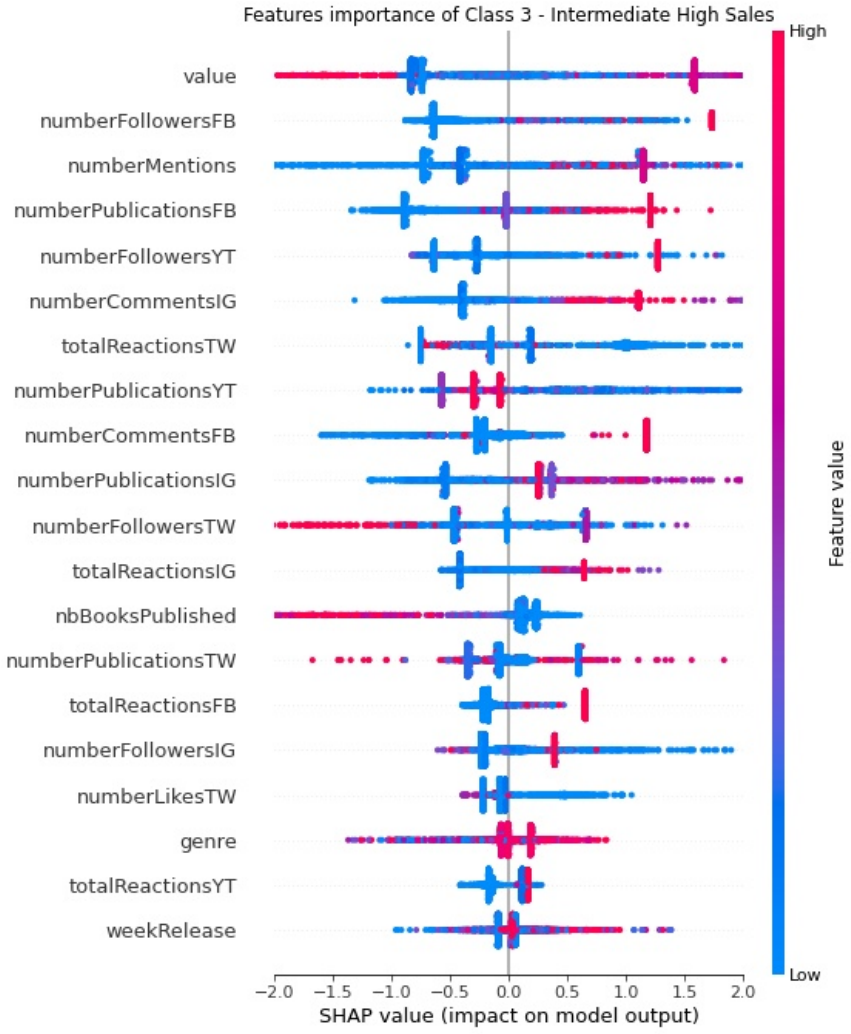

FIGURE A.9: Summary Plot to see the relationship of all the variables with the model for Class 3, and the impact they have. Source: Own elaboration.

Finally, Figure A.10 the value -9.94 is the prediction of the model for Class 4 according to the characteristics that were input to it. Here it can be seen that the most relevant variable is `nbBooksPublished`, although others can be appreciated such as: `value`, `numberCommentsYT`. The impact of these variables in this category are analyzed in more detail below.

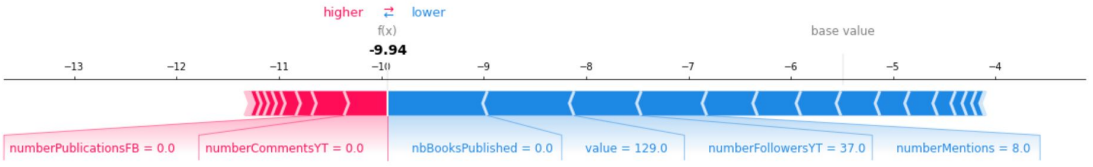

FIGURE A.10: Force Plot to visualize the importance of each input parameter in the model prediction for Class 4. Source: Own elaboration.

In Figure A.11, belonging to Class 4 - High sales, it can be seen that the feature `nbBooksPublished`, it is mostly associated with low and positive values with respect to sales. This implies that people buy more books if the number of books that the author has previously published is low. Regarding the characteristic `value`, again there is a variety, since the vast majority of people will buy more books, if they are affordable, and another part that will buy more books, if they are priced affordable high. These values indicate a similarity with the class to which they are associated (High Sales = Bestseller). On the other hand, the feature `yearRelease` is the least important, indicating that the vast majority of people will buy fewer books if the release year is more current.

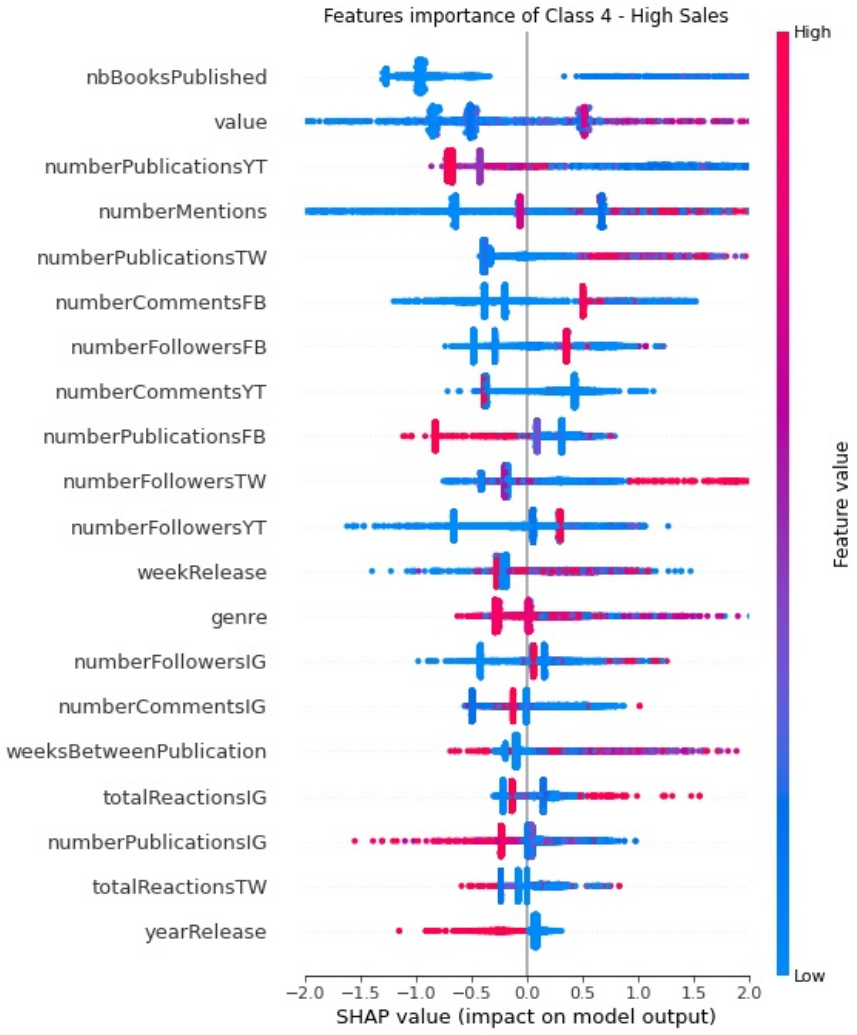

FIGURE A.11: Summary Plot to see the relationship of all the variables with the model for Class 4, and the impact they have. Source: Own elaboration.

Summary

The data show that, evidently, the segmentation by the experts is basically governed by the financial variable; since in all classes, the feature `value` is reflected among the most relevant variables. Other characteristics that globally influence the purchase of books is the use of new technologies such as social networks (Facebook, Instagram, YouTube).

A.0.3 Quartiles segmentation classes analysis

In Figure A.12 the value 6.29 is the prediction of the model for Class 1 according to the characteristics that were input to it. The parameters that are in red are the ones that make the prediction have a higher value, while the parameters that are in blue make the prediction have a lower value. It can be clearly seen that the variable `value` has high relevance, which causes the prediction to increase in value from the base value; followed by `numberFollowersTW`, `numberLikesTW`. The impact of these variables in this category are analyzed in more detail below.

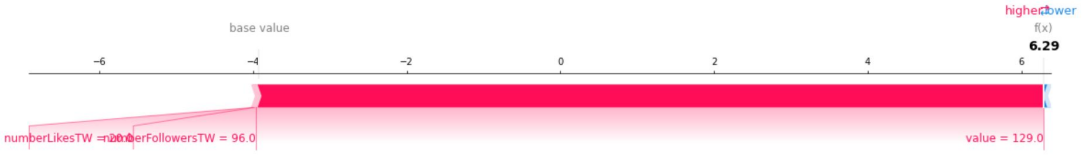

FIGURE A.12: Force Plot to visualize the importance of each input parameter in the model prediction for Class1-Low sales. Source: Own elaboration.

In Figure A.13, belonging to Class 1 - Low sales, it can be seen that the features are ordered by Tree Shape and, therefore, the only thing that influences sales (for this class) is the variable `value` it is mostly associated with high and low values and negative with respect to sales. Implying that people buy more books if the price is low, given that, as can be seen, the pink color distribution stands out more than the blue ones. The rest of the variables do not show any impact on sales.

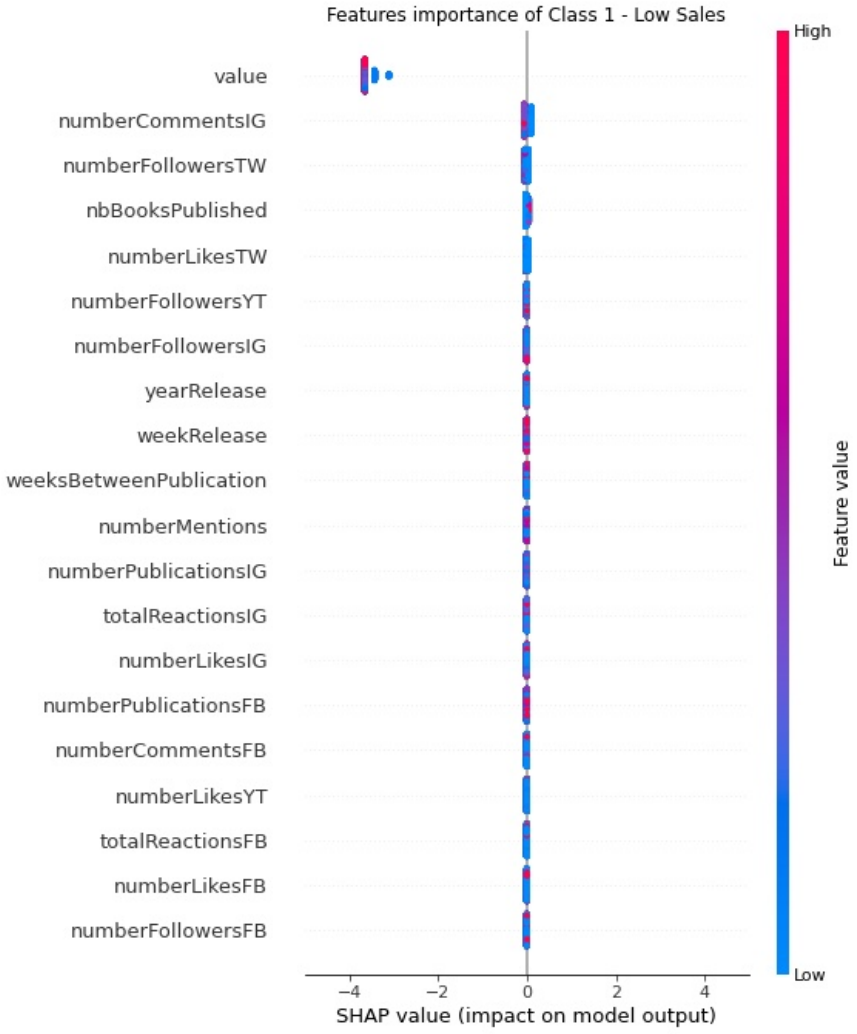

FIGURE A.13: Summary Plot to see the relationship of all the variables with the model for Class1, and the impact they have. Source: Own elaboration.

If Class 2 - Intermediate Low Sales is analyzed, it can be seen in Figure A.14 the most relevant variables are `value` and `weekBetweenPublication` and `genre`. The impact of these variables in this category are analyzed in more detail below.

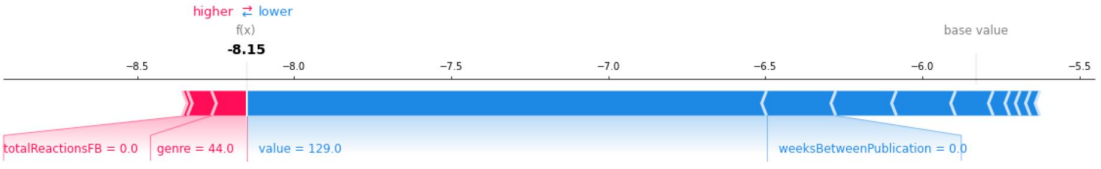

FIGURE A.14: Force Plot to visualize the importance of each input parameter in the model prediction for Class 2. Source: Own elaboration.

In Figure A.15, belonging to Class 2 - Intermediate low sales, it can be seen that the feature `value` presents a bipolarity of opinions on the part of people

who buy books. On the one hand, they will buy fewer books if the price is high, but also those books that are priced too low. If we analyze the variable `weeksBetweenPublication`, it indicates that people will buy more books if there is a greater publication margin, for example: between a first book and a second by an author. The same is reflected in the variable below `weekRelease` indicates that people will buy more books if there are fewer books released by the same author. From the analysis of the following two characteristics, we can intuit that customers will buy more books if there is less interaction on the Twitter and YouTube platforms.

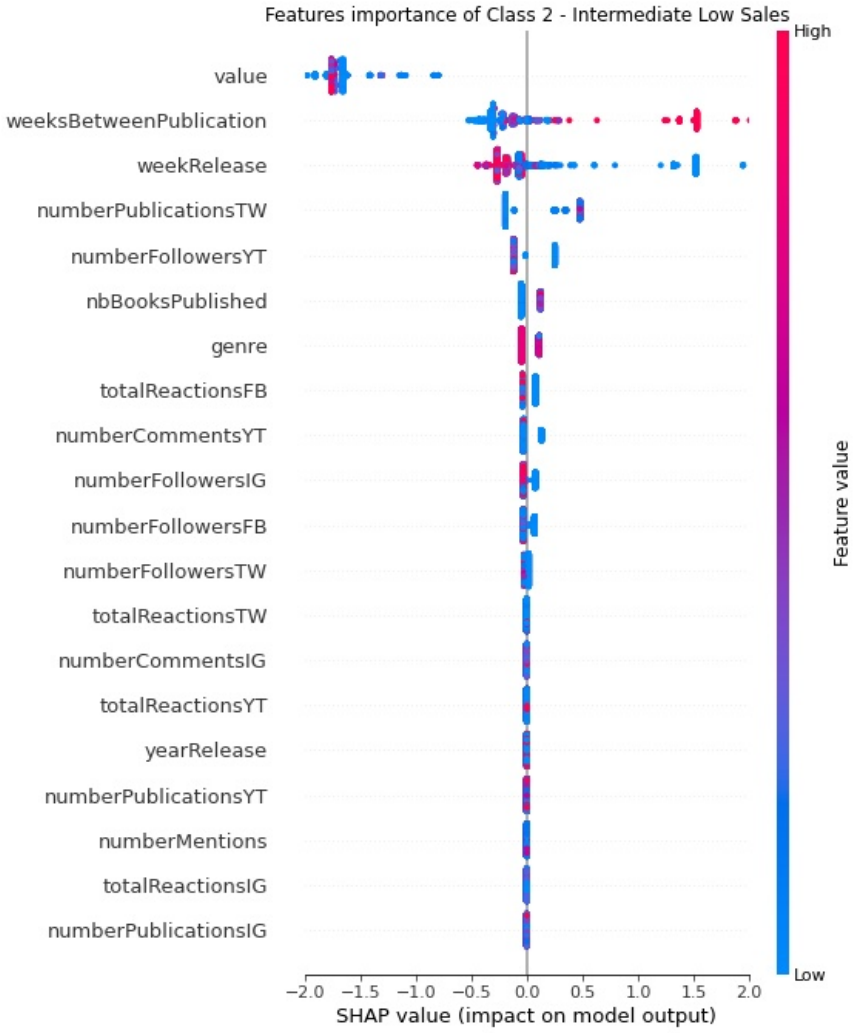

FIGURE A.15: Summary Plot to see the relationship of all the variables with the model for Class 2, and the impact they have. Source: Own elaboration.

In Figure A.16 the features more relavant are `value`, `totalReactionsYT` and `numberCommentsYT`. The impact of these variables in this category are analyzed in more detail below.

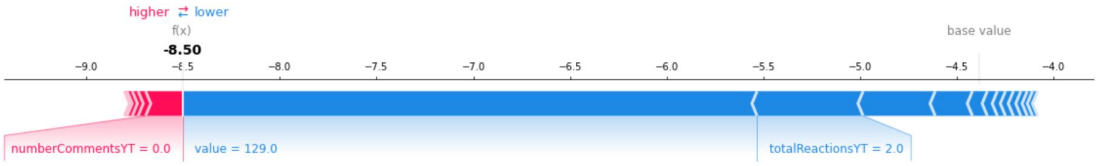

FIGURE A.16: Force Plot to visualize the importance of each input parameter in the model prediction for Class 3. Source: Own elaboration.

In Figure A.17, belonging to Class 3 - High intermediate sales, it can be seen that clearly people do not buy books that belong to this group, if they have a very high price. However, they will buy more if they are promoted on platforms such as YouTube (due to the reactions) and Facebook (the number of followers), this indicator will be interesting so that the commercial department is aware of which platforms to use for their advertising campaigns. On the other hand, on other platforms such as Instagram or Twitter, they do not seem to influence this purchase decision as much.

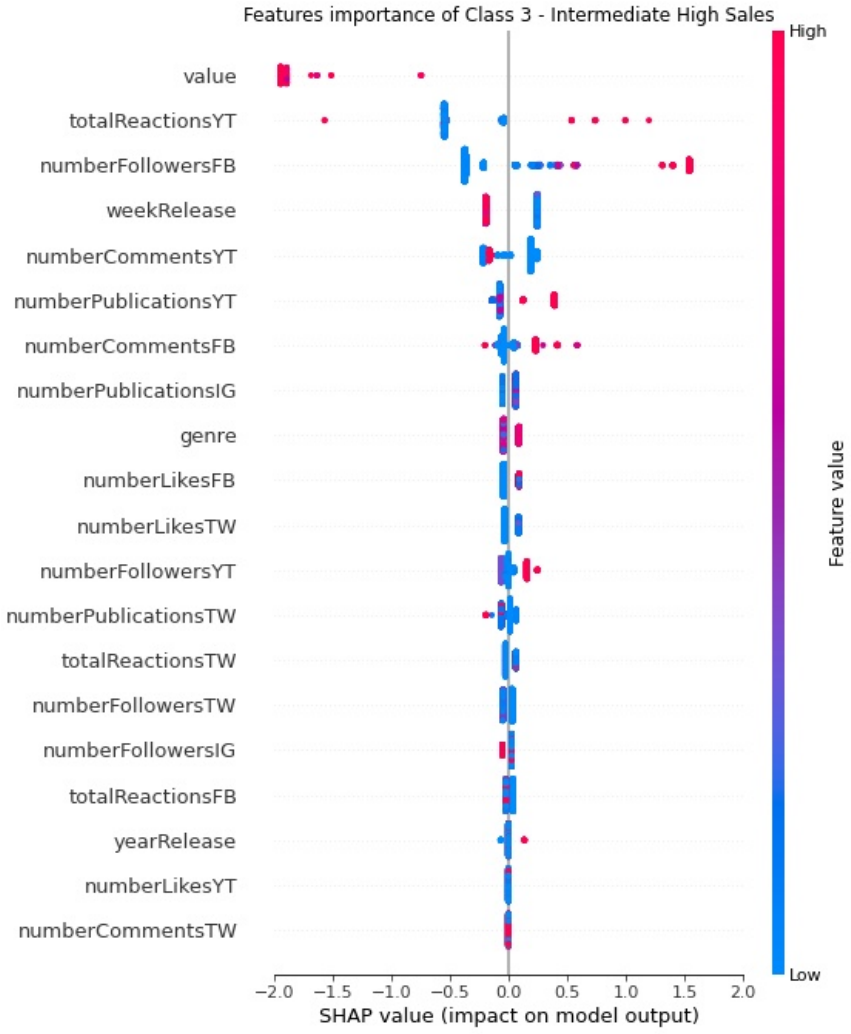

FIGURE A.17: Summary Plot to see the relationship of all the variables with the model for Class 3, and the impact they have. Source: Own elaboration.

Finally, Figure A.18 the value -7.45 is the prediction of the model for Class 4 according to the features that were input to it. Here it can be seen that the most relevant variables are `value` and `weekRelease`. The impact of these variables in this category are analyzed in more detail below.

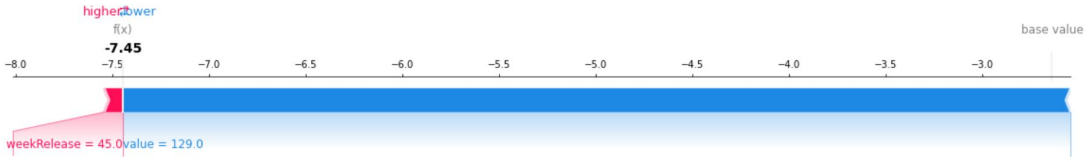

FIGURE A.18: Force Plot to visualize the importance of each input parameter in the model prediction for Class 4. Source: Own elaboration.

In Figure A.19, belonging to Class 4 - High sales, it can be seen that customers who buy books in this category will not buy if the price is too low, it is also appreciated

that the fact that new book releases are delayed over time will influence them to buy more books. This is a good indicator for publishers of "Bestseller" books, not to publish one after the other, but to provide a margin between them. The rest of the variables do not seem to influence the purchase decision too much.

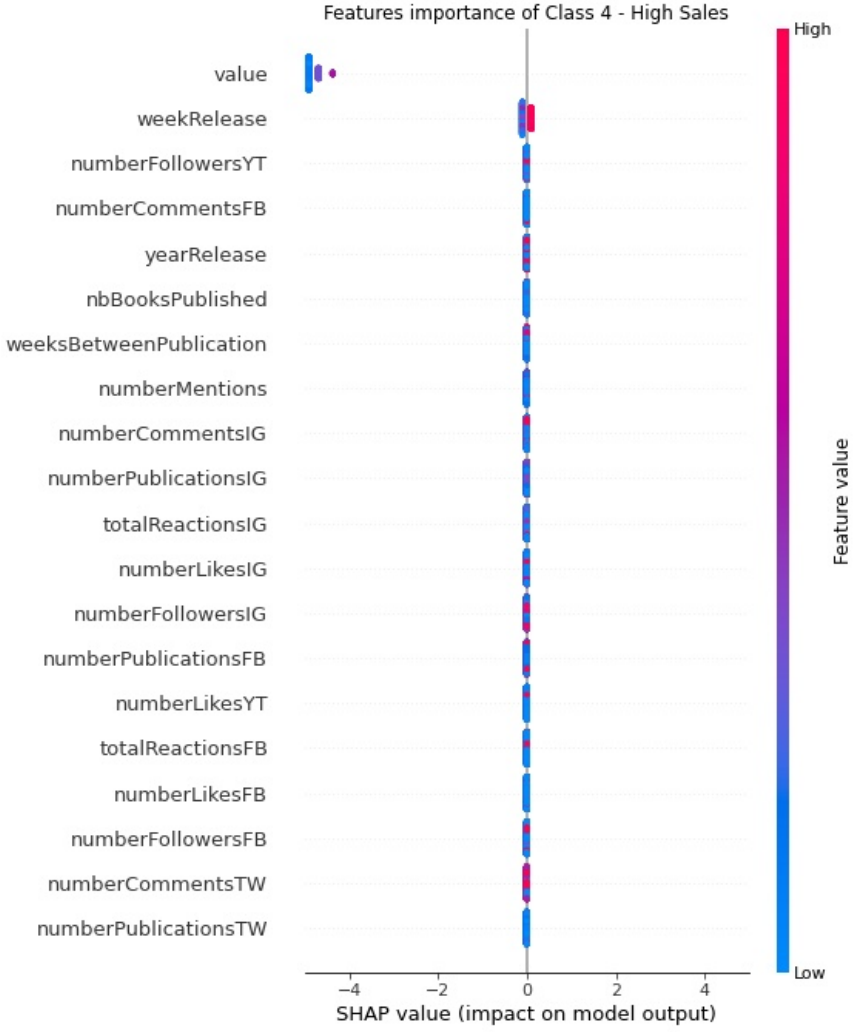

FIGURE A.19: Summary Plot to see the relationship of all the variables with the model for Class 4, and the impact they have. Source: Own elaboration.

Summary

As has happened with the segmentation of the experts, in the segmentation by quartiles the variable that most influences each of the classes is the price, rather than by the analysis of the behavior of the data. Unlike the segmentation by experts (and being consistent with the segmentation carried out), the use of social networks does not influence the purchase of a book as much. Also in part, because the data set has been divided into 4 parts, without taking into account any specific variable; as if the classification given by the experts can do it. Although the segmentation by quartiles is not influenced by any specific variable, it does not reflect the real behavior that the data may present, as in the segmentation by clusters.

A.0.4 Predictive results after applying the regression model

Figure A.20 shows a sample of the predictive results after applying the regression model to the entire data set. To the test data set, the real value of the number of copies of the book has been added, in order to verify if the predicted value is close or not, an error margin column has also been added, to calculate the error between these values.

AUTHOR	TITLE	GENRE3	YEAR_RE	WEEK_RI	NBBOOKSPUBI	WEEKSBETWEENPUBI	18-24	MENTIONSNB	FNI	ERI	UNITSACCUM	UNITSPREDICTED	ERRORMARGIN
ADEXE & NAU	TU Y YO. ABRAZANDO UN SUEÃ	LITERATURA J	2018	7	0.0	0.0	221	1116.0	1154972.0	0.0324	7292.0	6972.0	-320.0
ALAYRACH, MIGUEL	MI CAMA ES MUY ELAAAAAAAS	LITERATURA I	2018	45	0.0	0.0	0	8.0	457.0	-2.0	129.0	254.0	125.0
ALAYRACH, MIGUEL	BLANQUITO NO PINTA NADA	LITERATURA I	2018	48	1.0	3.0	0	8.0	457.0	-2.0	250.0	100.0	-150.0
ALBERT, MELISSA	THE HAZEL WOOD	LITERATURA J	2018	5	0.0	0.0	1	21.0	9055.0	0.0713	18.0	128.0	110.0
ARTIGAU I QUERALT, I	LES PARAULES INUTILS	LITERATURA J	2019	35	1.0	81.0	38	283.0	1965.0	-2.0	482.0	822.0	340.0
BALMES, SANTI	JO ANIRE A L'ESCOLA PER TU	LITERATURA I	2018	45	0.0	0.0	538	2678.0	100479.0	0.048	972.0	1126.0	154.0
CADEPE, MARIA	OTRO NIVEL, EL	LIBRO DIDACT	2018	7	0.0	0.0	0	0.0	180973.0	0.0097	3901.0	2251.0	-1650.0
CAMBLOR, GEMMA	MACETA ENCANTADA, LA	LITERATURA I	2019	14	0.0	0.0	6	24.0	11035.0	0.033	2408.0	3108.0	700.0
PAPA FRANCISCO	PAPA FRANCISCO DIJO NO, EL.	RELIGION	2018	8	0.0	0.0	1336	9379.0	7673493.0	0.0203	228.0	280.0	52.0
PAPA FRANCISCO	DIOS ES JOVEN	RELIGION	2018	12	4.0	1.0	1336	9379.0	7673493.0	0.0203	3686.0	2361.0	-1325.0
USART, MONICA	EL TEMPS A LA MALETA	GEOLOGIA	2019	10	0.0	0.0	0	0.0	33294.0	0.0481	1162.0	1442.0	280.0
ABAD CADENAS, CRIST	LIBERTAD DE AMAR, LA. GUAD	MEMORIAS Y	2018	42	0.0	0.0	2	14.0	-1.0	-1.0	4888.0	3958.0	-930.0
ADRIA ACOSTA, ALBEF	TICKETS EVOLUTION	GASTRONOM	2018	12	0.0	0.0	0	0.0	138214.0	-2.0	3532.0	3562.0	30.0
AGUILAR CARRASCO, I	PAPEL DE LAS MUJERES EN EL C	SOCIOLOGIA	2018	3	0.0	0.0	483	2951.0	799.0	-2.0	190.0	197.0	7.0
AGUILAR CARRASCO, I	FEMINISMO O BARBARIE (VOL.	SOCIOLOGIA	2018	20	1.0	17.0	483	2951.0	799.0	-2.0	149.0	59.0	-90.0
AMORETTI, ANDREA	EMPIEZA POR LOS ZAPATOS. CC	AUTOAYUDA	2018	11	0.0	0.0	0	0.0	30007.0	0.0196	1950.0	2815.0	865.0

FIGURE A.20: Predictive results after applying a regressor model on the entire test data set. Source: Own elaboration.

Figure A.21 shows the predictive results applied to the same sample but applying the combined model created, which initially places the books in one of the 4 classes and later calculates the regression of each of these classes. As the values are observed despite presenting errors, they are able to predict with greater accuracy than applying a global regressor.

AUTHOR	TITLE	GENRE3	YEAR_RE	WEEK_RI	NBBOOKSPUBL	WEEKSBETWEENPUBI	18-24	MENTIONSNB	FNI	ERI	CLASS	UNITSACCUMULATETOT/	UNITSPREDICTED	ERRORMARGIN
ADEXE & NAU	TU Y YO. ABRAZANDO UN SUEÃ	LITERATURA J	2018	7	0.0	0.0	221	1116.0	1154972.0	0.0324	3.0	7292.0	7292.0	0.0
ALAYRACH, MIGUEL	MI CAMA ES MUY ELAAAAAAAS	LITERATURA I	2018	45	0.0	0.0	0	8.0	457.0	-2.0	1.0	129.0	219.0	90.0
ALAYRACH, MIGUEL	BLANQUITO NO PINTA NADA	LITERATURA I	2018	48	1.0	3.0	0	8.0	457.0	-2.0	1.0	250.0	165.0	-85.0
ALBERT, MELISSA	THE HAZEL WOOD	LITERATURA J	2018	5	0.0	0.0	1	21.0	9055.0	0.0713	3.0	18.0	18.0	0.0
ARTIGAU I QUERALT, I	LES PARAULES INUTILS	LITERATURA J	2019	35	1.0	81.0	38	283.0	1965.0	-2.0	4.0	482.0	495.0	13.0
BALMES, SANTI	JO ANIRE A L'ESCOLA PER TU	LITERATURA I	2018	45	0.0	0.0	538	2678.0	100479.0	0.048	4.0	972.0	1066.0	94.0
CADEPE, MARIA	OTRO NIVEL, EL	LIBRO DIDACT	2018	7	0.0	0.0	0	0.0	180973.0	0.0097	2.0	3901.0	3894.0	-7.0
CAMBLOR, GEMMA	MACETA ENCANTADA, LA	LITERATURA I	2019	14	0.0	0.0	6	24.0	11035.0	0.033	2.0	2408.0	2416.0	8.0
PAPA FRANCISCO	PAPA FRANCISCO DIJO NO, EL I	RELIGION	2018	8	0.0	0.0	1336	9379.0	7673493.0	0.0203	4.0	228.0	280.0	52.0
PAPA FRANCISCO	DIOS ES JOVEN	RELIGION	2018	12	4.0	1.0	1336	9379.0	7673493.0	0.0203	4.0	3686.0	3586.0	-100.0
USART, MONICA	EL TEMPS A LA MALETA	GEOLOGIA	2019	10	0.0	0.0	0	0.0	33294.0	0.0481	3.0	1162.0	1162.0	0.0
ABAD CADENAS, CRIST	LIBERTAD DE AMAR, LA. GUADA	MEMORIAS Y	2018	42	0.0	0.0	2	14.0	-1.0	-1.0	3.0	4888.0	4888.0	0.0
ADRIA ACOSTA, ALBER	TICKETS EVOLUTION	GASTRONOM	2018	12	0.0	0.0	0	0.0	138214.0	-2.0	2.0	3532.0	3538.0	6.0
AGUILAR CARRASCO, I	PAPEL DE LAS MUJERES EN EL C	SOCIOLOGIA	2018	3	0.0	0.0	483	2951.0	799.0	-2.0	1.0	190.0	197.0	7.0
AGUILAR CARRASCO, I	FEMINISMO O BARBARIE (VOL.	SOCIOLOGIA	2018	20	1.0	17.0	483	2951.0	799.0	-2.0	1.0	149.0	99.0	-50.0
AMORETTI, ANDREA	EMPIEZA POR LOS ZAPATOS. CC	AUTOAYUDA	2018	11	0.0	0.0	0	0.0	30007.0	0.0196	2.0	1950.0	1942.0	-8.0

FIGURE A.21: Predictive results after applying a regressor model on the entire test data set. Source: Own elaboration.

A.0.5 Predicted number of book copies by new authors

The following visualization A.22 shows the number of book copies that the system predicts for each new author. These are examples given and validated by "The Editorial". Whose authors are mostly influencers or public figures, of which the publisher was interested in knowing the number of copies they could sell before the launch of their books. This prediction is made during the year 2020.

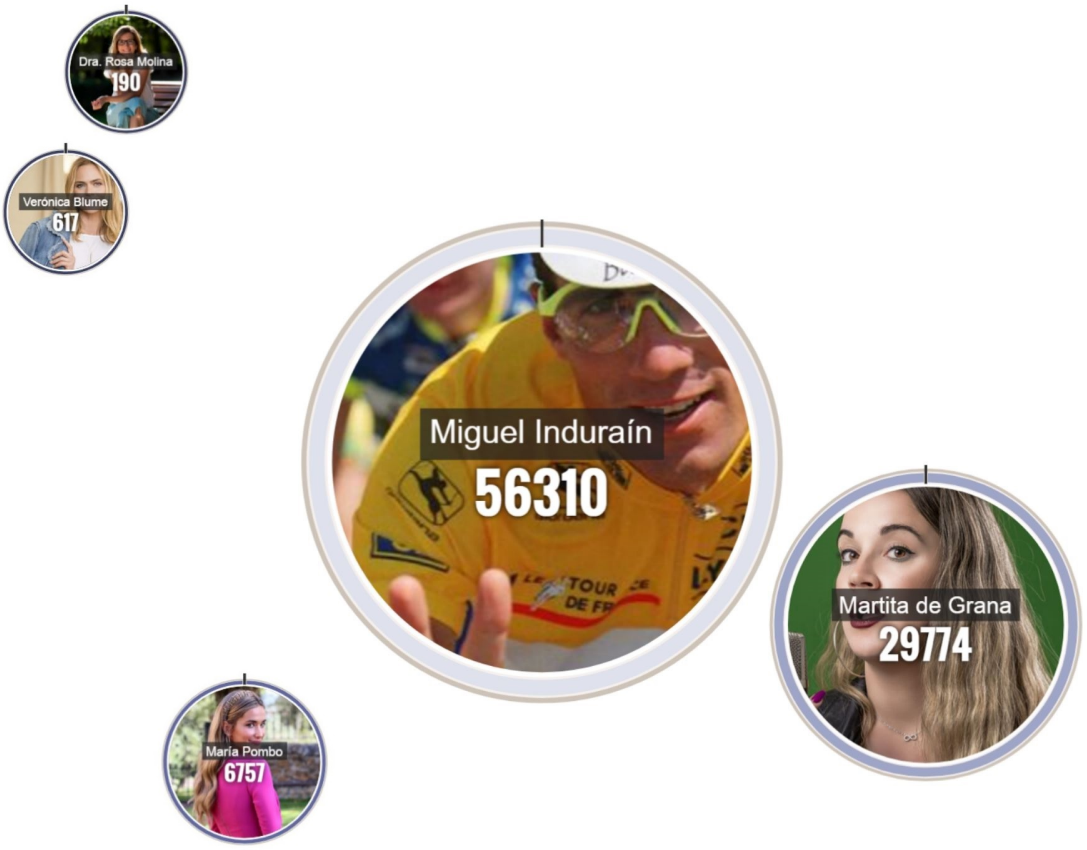

FIGURE A.22: Predicted number of book copies by new authors. Source: Own elaboration adapted from: https://www.diariodesevilla.es/2021/06/09/entrevistas/Martita-Grana_1581752207_139766695_667x375.jpg; https://www.semana.es/wp-content/uploads/2019/10/maria-pombo-cumpleanos-dest.jpg; https://www.clara.es/medio/2019/11/18/veronica-blume-clara-revista-mes-mayo_3317cae0_1280x772.jpg; https://www.bizkaiatalent.eus/wp-content/uploads/2017/03/miguel-indurain-430x260.jpg and https://www.elindependiente.com/wp-content/uploads/2021/10/ENK4115-scaled.jpg

Appendix B

Personality Profile

B.0.1 Entity recognition tests

This section shows the complete example of the entity recognition tests of a paragraph from the work "The prodigious afternoon of Baltazar", both with the pretrained model Figure B.1, and the fitted model Figure B.2.

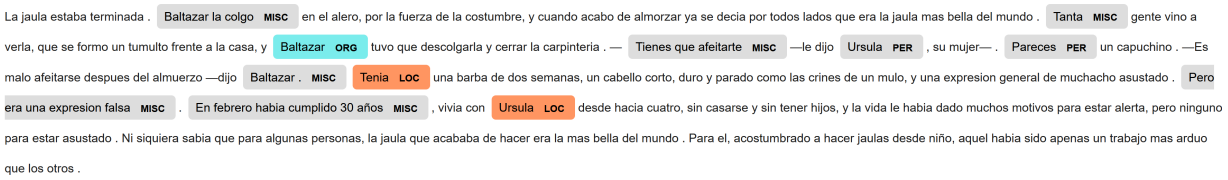

FIGURE B.1: NER with pretrainned model. Source: Own elaboration.

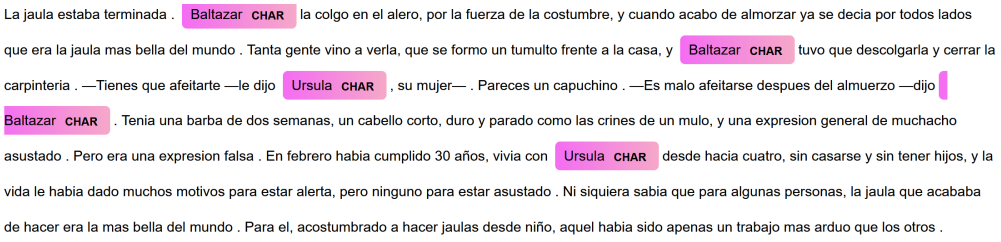

FIGURE B.2: NER with my model. Source: Own elaboration.

B.0.2 Character recognition distributions

This section shows in Figure B.3 the character recognition distributions of the remaining works, after comparing the pretrained model and my model.

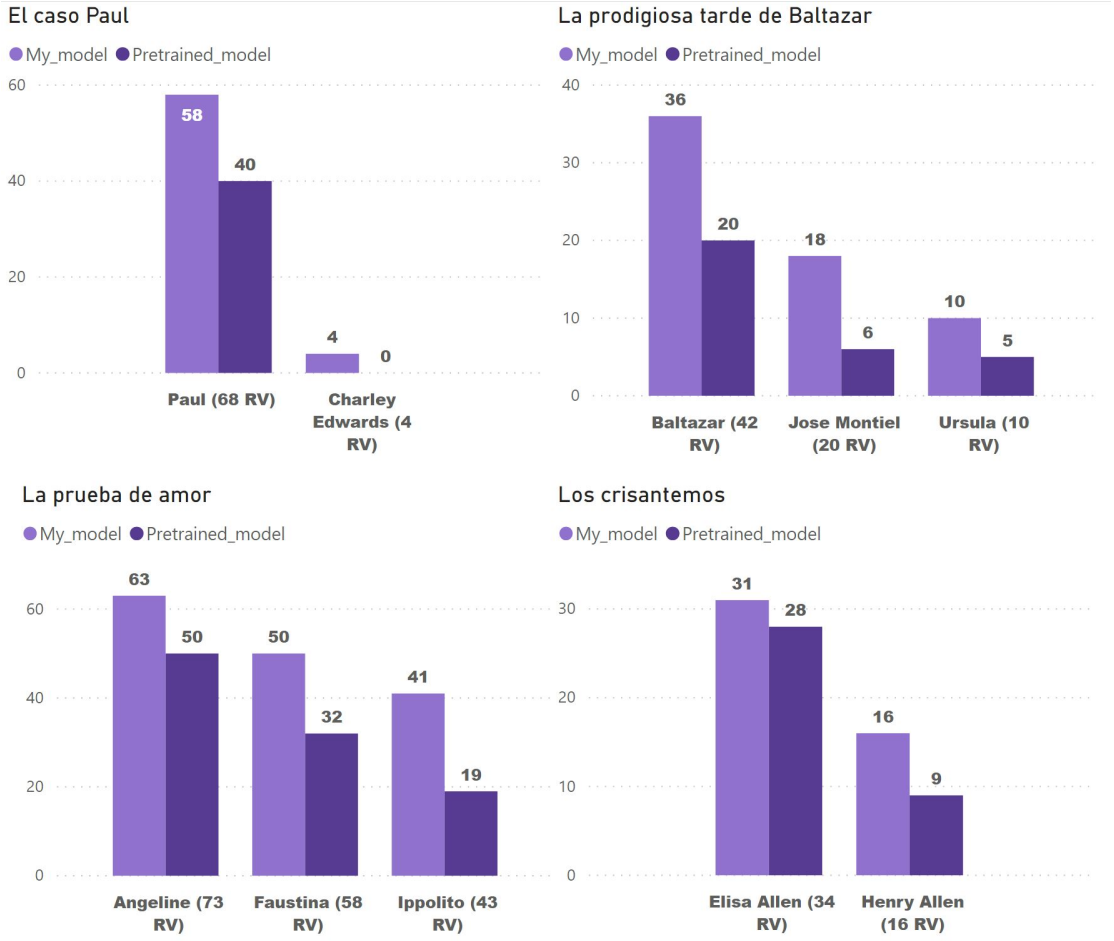

FIGURE B.3: Result of characters found with the pretrained model of spaCy versus model fitted with the literary theme. The real number of appearances of character in the work on the X axis next to the name of character. Making a comparison between the models, it can be seen that the results of the adjusted model are much more accurate than those of the pretrained model. Source: Own elaboration.

Bibliography

[1] Vanguardia2021. Los críticos literarios recomiendan. https://www.lavanguardia.com/cultura/culturas/20211218/7933726/libros-regalo-navidad-recomendacion.html. Accessed: 2023-04-11.

[2] MarketingNews2023. La campaña de este año se centra en el papel del libro como elemento de unión. https://www.marketingnews.es/marcas/noticia/1165472054305/campana-de-ano-se-centra-papel-del-libro-elemento-de-union-marc-rocamora-p1.html#. Accessed: 2023-04-11.

[3] Pais1998. Libros: cantidad y calidad. https://elpais.com/diario/1998/05/24/opinion/895960808_850215.html. Accessed: 2023-04-11.

[4] Inbouncycle2022. Marketing editorial: estrategias digitales en la industria del libro. https://www.inboundcycle.com/blog-de-inbound-marketing/marketing-editorial. Accessed: 2023-04-11.

[5] I Megías Queirós and Elena Rodríguez San Julián. Jóvenes en el mundo virtual: usos, prácticas y riesgos. *Madrid: Centro Reina Sofía sobre Adolescencia y Juventud. FAD. Fundación Mapfre*, 2018.

[6] Jessie Caridad Martín Sujo, Elisabet Golobardes i Ribé, Xavier Vilasís Cardona, Virginia Jiménez Ruano, and Javier Villasmil López. Correction to: Smartdata: An intelligent decision support system to predict the readers permanence in news. In *Proceedings of SAI Intelligent Systems Conference*, pages C1–C1. Springer, 2022.

[7] Miroslav Kubat, Ivan Bratko, and Ryszard S Michalski. A review of machine learning methods. *Machine learning and data mining: methods and applications*, pages 3–69, 1998.

[8] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[9] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[10] Mohamed Farouk Abdel Hady and Friedhelm Schwenker. Semi-supervised learning. *Handbook on Neural Information Processing*, pages 215–239, 2013.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[12] Vlado Keselj. Speech and language processing daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, xxxi+ 988 pp; hardbound, isbn 978-0-13-187321-6, $115.00, 2009.

[13] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[14] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*, 1802.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Paul Resnik and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[22] Nathaniel Good, J Ben Schafer, Joseph A Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, John Riedl, et al. Combining collaborative filtering with personal agents for better recommendations. *Aaai/iaai*, 439(10.5555): 315149–315352, 1999.

[23] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[24] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201, 1995.

[25] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[26] Upendra Shardanand and Pattie Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 210–217, 1995.

[27] Collaborative Filtering based Recommendation Systems exemplified. https://towardsdatascience.com/collaborative-filtering-based-recommendation-systems-exemplified-ecbffe1c2 [Online; accessed May-2022].

[28] Robin Burke. The wasabi personal shopper: A case-based recommender system. In *AAAI/IAAI*, pages 844–849, 1999.

[29] Robin Burke. Knowledge-based recommender systems. *Encyclopedia of library and information systems*, 69(Supplement 32):175–186, 2000.

[30] Sascha Schmitt and Ralph Bergmann. Applying case-based reasoning technology for product selection and customization in electronic commerce environments. In *12th bled electronic commerce conference*, volume 273, 1999.

[31] Jerome H Friedman, Forest Baskett, and Leonard J Shustek. An algorithm for finding nearest neighbors. *IEEE Transactions on computers*, 100(10): 1000–1006, 1975.

[32] Keinosuke Fukunaga and Patrenahalli M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753, 1975.

[33] The Myers-Briggs Foundation. https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/original-research.htm. [Online; accessed May-2022].

[34] Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

[35] J. Park, G.L. Ciampaglia, and E. Ferrara. Style in the age of instagram: Predicting success within the fashion industry using social media. *In Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing, San Francisco, CA, USA*, pages 64–73, 27 February–2 March 2016.

[36] N.B. Lassen, R. Madsen, and R. Vatrapu. Predicting iphone sales from iphone tweet. *In Proceedings of the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, Ulm, Germany*, pages 81–90, September 2014.

[37] F. Abel, E. Diaz-Aviles, N. Henze, D. Krause, and P. Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. *In Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark*, pages 276–280, August 2010.

[38] G.C. Moon, G. Kikuta, T. Yamada, A. Yoshikawa, and T. Terano. Blog information considered useful for book sales prediction. *In Proceedings of the 7th International Conference on Service Systems and Service Management, Tokyo, Japan*, pages 1–5, June 2010.

[39] A. Rapp, L.S. Beitelspacher, D. Grewal, and D.E. Hughes. Understanding social media effects across seller, retailer, and consumer interactions. *Journal of the Academy of Marketing Science*, pages 547–566, 2013.

[40] R. Guesalaga. The use of social media in sales: Individual and organizational antecedents, and the role of customer engagement in social media. *Industrial Marketing Management*, 54:71–79, 2016.

[41] X. Wang, B. Yucesoy, O. Varol, T. Eliassi-Rad, and A. L. Barabási. Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1): 1–20, 2019.

[42] N. Kim and W. Kim. Do your social media lead you to make social deal purchases? consumer-generated social referrals for sales via social commerce. *International Journal of Information Management*, 39:38–48, 2018.

[43] B. Yucesoy, X. Wang, J. Huang, and A. L. Barabási. Success in books: a big data approach to bestsellers. *EPJ Data Science*, 7:1–25, 2018.

[44] T. Q. Feng, M. Choy, and M. N. Laik. Predicting book sales trend using deep learning framework. *International Journal of Advanced Computer Science and Applications*, 11(2):28–39, 2020.

[45] H. Rew. Francis galton. *JR Stat*, 85:293–298, 1922.

[46] W. Winfrey and L. Heaton. Market segmentation nalysis of the indonesian family planning market: Consumer, provider and product market segments and public sector procurement costs of family planning under. *USAID*, 1996.

[47] H. Lehmann and A. Zaiceva. Informal employment in russia: Incidence, determinants and labor market segmentation, tipo. https://ssrn.com/abstract=2330214. Accessed: 2021-01-15.

[48] G.T. Duncan, W.L. Gorr, and J. Szczypula. Forecasting analogous time series. *In Principles of Forecasting Springer: Boston, MA, USA*, pages 195–213, 2001.

[49] Elizabeth A Maharaj and Brett A Inder. Forecasting time series from clusters. 1999.

[50] Richard Mitchell. Forecasting electricity demand using clustering. In *Applied Informatics*, volume 378, pages 378–134, 2003.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. ... Gomez, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[52] D. Vala, H.and Jurgens, A. Piper, and D. Ruths. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *In Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 769–774, 2015, September.

[53] X. Wei, H. Zang, and Q. Zhang. Disambiguate chinese personal pronoun based on semantic structure. In *2008 IEEE International Conference on Granular Computing*, pages 644–648, 2008, August.

[54] Y. H. Chen and J. D. Choi. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, 2016, September.

[55] W. C. Tirre and S. Dixit. Reading interests: Their dimensionality and correlation with personality and cognitive factors. personality and individual differences. 18(6):731–738, 1995.

[56] L. Flekova and I. Gurevych. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, 2015, September.

[57] D. K. Elson, K. McKeown, and N. J. Dames. Extracting social networks from literary fiction. 2010.

[58] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pages 1–16, 2005, June.

[59] J. Oberlander and S. Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634, 2006, July.

[60] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

[61] P. Cantos. El inglés y el español desde una perspectiva cuantitativa y distributiva: equivalencias y contrastes1/english and spanish from a distributional and quantitative perspective: Equivalences and contrasts. *Estudios ingleses de la Universidad Complutense*, 19:15–44, 2011.

[62] Roger C Schank and Robert Abelson. P.(1977) scripts, plans, goals and understanding: An inquiry into human knowledge structures, 1977.

[63] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1): 39–59, 1994.

[64] Roger C Schank. *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press, 1983.

[65] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[66] Ren Yu, Benoit Iung, and Hervé Panetto. A multi-agents based e-maintenance system with case-based reasoning decision support. *Engineering applications of artificial intelligence*, 16(4):321–333, 2003.

[67] Chen-Shu Wang and Heng-Li Yang. A recommender mechanism based on case-based reasoning. *Expert Systems with Applications*, 39(4):4335–4343, 2012.

[68] Delu Wang, Kaidi Wan, and Wenxiao Ma. Emergency decision-making model of environmental emergencies based on case-based reasoning method. *Journal of environmental management*, 262:110382, 2020.

[69] Parfait Atchade Adelomou, Daniel Casado Fauli, Elisabet Golobardes Ribé, and Xavier Vilasis-Cardona. Quantum case-based reasoning (qcbr). *Artificial Intelligence Review*, pages 1–27, 2022.

[70] Pei-Chann Chang and Chien-Yuan Lai. A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting. *Expert Systems with Applications*, 29(1):183–192, 2005.

[71] Yi-Fen Chen. Herd behavior in purchasing books online. *Computers in Human Behavior*, 24(5):1977–1992, 2008.

[72] Pei-Chann Chang, Chien-Yuan Lai, and K Robert Lai. A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting. *Decision Support Systems*, 42(3):1715–1729, 2006.

[73] Haitao Wu, Botao Zhong, Benachir Medjdoub, Xuejiao Xing, and Li Jiao. An ontological metro accident case retrieval using cbr and nlp. *Applied Sciences*, 10(15):5298, 2020.

[74] F Salager-Meyer. A text-type based discourse analysis of medical english. *Abstracts internal structuring*, 1991.

[75] Wael H Gomaa, Aly A Fahmy, et al. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18, 2013.

[76] Tomás López-Solaz, José A Troyano, F Javier Ortega, and Fernando Enríquez. Una aproximaciónn al uso de word embeddings en una tarea de similitud de textos en español. *Procesamiento del Lenguaje Natural*, (57): 67–74, 2016.

[77] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019.

[78] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Zero-shot fact verification by claim generation. *arXiv preprint arXiv:2105.14682*, 2021.

[79] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[80] Jacob Benesty, Jingdong Chen, and Yiteng Huang. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):757–765, 2008.

[81] Jessie C Martín Sujo, Elisabet Golobardes i Ribé, and Xavier Vilasís Cardona. Cait: A predictive tool for supporting the book market operation using social networks. *Applied Sciences*, 12(1):366, 2021.

[82] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[83] Shap Library. https://pypi.org/project/shap/. [Online; accessed Nov-2018].

[84] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[85] J Ross Quinlan. Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2):339–346, 1990.

[86] Decision Trees. https://scikit-learn.org/stable/modules/tree.html. [Online; accessed Dic-2019].

[87] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[88] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[89] Sklearn.ensemble.RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Online; accessed Dic-2019].

[90] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[91] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer, 2003.

[92] sklearn.neighbors.KNeighborsClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Online; accessed Dic-2019].

[93] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[94] Didrik Nielsen. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU, 2016.

[95] XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/. [Online; accessed Dic-2019].

[96] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[97] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[98] sklearn.ensemble.GradientBoostingClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html. [Online; accessed Jan-2020].

[99] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[100] sklearn.ensemble.GradientBoostingClassifier. https://lightgbm.readthedocs.io/en/latest/Python-Intro.html. [Online; accessed Jan-2020].

[101] Frank M Bass. A new product growth for model consumer durables. *Management science*, 15(5):215–227, 1969.

[102] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[103] Peter Bloomfield. An exponential model for the spectrum of a scalar time series. *Biometrika*, 60(2):217–226, 1973.

[104] RD Berger. Comparison of the gompertz and logistic equations to describe plant disease progress. *Phytopathology*, 71(7):716–719, 1981.

[105] V. I. Propp. *Morphology of the Folktale*. University of Texas Press, 9 edition, 1968.

[106] Myers-Briggs Personality Type Dataset. https://www.kaggle.com/datasets/datasnaek/mbti-type. [Online; accessed May-2022].

[107] Jessie C Martín Sujo and Elisabet Golobardes i Ribé. Personality profile of fictional characters in books using natural language processing. *Personality Profile of Fictional Characters in Books Using Natural Language Processing*.

[108] M. Honnibal and I. Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

[109] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, page ., 2018.

[110] Marvin L. Minsky and Seymour A. Papert. *Perceptrons: expanded edition*. MIT press, 1988.

[111] S. Haykin and R. Lippmann. Neural networks, a comprehensive foundation. *International journal of neural systems*, 5(4):363–364, 1994.

[112] David Keirsey. *Please understand me II: temperament, character, intelligence*. Prometheus Nemesis Book Company, 1998.

[113] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[114] Roberta finetuned. https://huggingface.co/titi7242229/roberta-base-bne-finetuned_personality_multi_2. [Online; accessed May-2022].

[115] Tweepy. https://www.tweepy.org/. [Online; accessed April-2022].

[116] MongoDB. https://www.mongodb.com/. [Online; accessed April-2022].

[117] Jessie Caridad Martín Sujo et al. Brain l: A book recommender system. *arXiv preprint arXiv:2302.00653*, 2023.

[118] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014.

[119] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[120] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. To appear.

[121] Transformers. https://pypi.org/project/transformers/. [Online; accessed Jun-2022].

[122] Gensim. https://pypi.org/project/gensim/. [Online; accessed Jun-2022].

[123] Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *arXiv preprint arXiv:1707.08172*, 2017.

[124] Agenda2030. https://www.agenda2030.gob.es/recursos/docs/METAS_DE_LOS_ODS.pdf". [Online; accessed Nov-2021].

[125] Sridhar Alla and Suman Kalyan Adari. What is mlops? In *Beginning MLOps with MLFlow*, pages 79–124. Springer, 2021.

[126] Aida Valls and Karina Gibert. Women in artificial intelligence. *Applied Sciences*, 12(19):9639, 2022.

[127] Aida Valls and Karina Gibert. *Women in Artificial intelligence (AI)*. MDPI, Basel, 2022. ISBN 978-3-0365-5532-4. doi: 10.3390/books978-3-0365-5532-4.