A CARS of Novels with Imbalanced Sample Treatment

Lucas de Faccio
lucasfaccionunes@gmail.com
Universidade de São Paulo
São Paulo, Brazil

Carla Marcolin
cbmarcolin@gmail.com
Universidade Federal de Uberlândia
Uberlândia, Brazil

Fábio Lobato
lobato.fabiof@gmail.com
Universidade Federal do Oeste Pará
Santarém, Brazil

ABSTRACT

Widely spread, recommender systems might face some challenges such as overspecialization and lack of diversity. In this paper, we propose a book context-aware recommender system (CARS) that uses individual characteristics as model features and active search as a pre-filtering context method in an attempt to increase user's newness perception and diversity. To achieve this goal, we revised literary critic essays to create five binary base-questions able to separate and aggregate novels through subjective concepts. We also conducted a data collection to form a dataset around 50 selected books, evaluated by the public using these questions. Going further, we developed two recommender systems (RS) using different strategies to handle imbalanced samples (SELC and SMOTE) and compare their performance to conclude that SELC generates better recommendations on an inner performance.

KEYWORDS

Context-aware recommender systems, Pre-filtering, Novels, Imbalanced sample, SMOTE

1 INTRODUCTION

Recommender systems (RS) are part of our lives, probably more than we see. They are in entertainment platforms, e-commerce websites and social media. However, RS development still face some challenges.

To improve the recommendations, RS should be more customizable, incorporating personalized aspect of the user to provide better suggestion, more suitable for their desire. This individual information is what researchers called *context*.

The context is a personalized piece of information about that user that would improve the RS so it would give a more accurate suggestion. For example, giving better item recommendation in an e-commerce website based in the purchase [1] or filtering songs from a ordinary playlist depending on the user's mood [2]. These recommender systems that use context to improve suggestions are called context-aware recommender system (CARS) (See Adomavicius and Tuzhilin [3] for more details).

Another relevant topic concerning recommendations is providing a way so the user explore content. The overspecialization of RS can create a bubble and submerge the user, creating a felling of lack of newness or diversity, sometimes called serendipitous recommendations [4]. Many strategies were posed to overcome this problem, including ontology language in order to achieve better knowledge about the recommended content; although, as Javed et al. [5] say, this approach might be challenging due to the complexity in developing ontologies.

Besides these, content-based RS as well as collaborative ones might suffer from the New Item and New User Problem considering these require a certain amount of ratings and interactions so that systems are able to recommend the items or to users properly [6].

Considering different challenges regarding RS, in this paper, we propose a CARS designed to recommend novels which uses two main mechanisms: **active search** and **pre-filtering context**, to let users decide what type of content they want and reduce the risk of overspecialization/serendipitous recommendation. The recommendations are made based on a pattern, specified by the user, that will be used to filter a database of novels before the application of a traditional recommendation algorithm.

The difference in the proposed algorithm is the way each novel is categorized. In order to have better accuracy, we tried to go further than similarities between item and users, and group novels based on their specifics. To do this we turned into literary and critic studies and searched for characteristics that could help us classify novels accordingly. The feature engineering in imbalanced data is a relevant problem, da Silva Mendes and de Jesus [7] studied this issue through machine learning models and feature selection to employee attrition as a study case.

From this review, we created five binary base-questions. Each question was designed to split the novels into two exclusive groups and deal with subjective concepts around the characters and the plot within the book. Independently of the story, author or publication period, any novel would fit in one side of the base-question. This set of questions is the foundation of the development and works as context for the items.

Figure 3 shows a flowchart on how the algorithm works. In the following sections, we explain each individual part, corresponding to the red numbers in the figure.

In section 2, we explain how each base-question was designed and what aspect is it dealing with. In the section 3, we present the data collection process we conducted considering the need of a dataset with books evaluated according to the base-questions. We also analyze some aspects around imbalanced samples. Next, in section 4, we present each step of the algorithm and introduce two solutions to overcome the imbalanced sample issue: one novel proposition designed by the authors and one classical technique. Then, throughout section 5 we evaluate the methods using binary classification performance measures. And finally, in section 6, we bring some conclusions and future developments to the work.

2 BASE-QUESTIONS

Probably, the most ancient and still relevant work about the dramatic arts, in Aristotle [8], the author analysed the Greek drama and theatrical literature scene. He defined the standard basis for a play to be classified as a top tragedy, in opposite with comedies; the former the most refined type of art and the latter considered a minor art branch. According to him, the epic poetry and the tragedy are equal, except by their rhythmic and narrative flow. Going further,

he established the tragedy has six main elements, focusing more in the more important ones, in order: plot and character.

Aristotle and the Poetics set most of the ground where the novels would born centuries later. And through time, authors explored the characteristics of tragedies, comedies and epics, which have built the foundations of the literary genre.

In what concerns to structure and theme, the modern novels are far from what they were in the early 18^{th} century, when the first European ones came up with Dom Quixote de La Mancha, by Miguel Cervantes, and Moll Flanders, by Daniel Defoe [9]. And they are even further from the topics Aristotle analysed in the Poetics. Even so, it can be seen a similarity between tragedy and novel, specially when we compare the way both of them do their main goal: to tell a story. In Oliveira [10], the author explains how narratives evolved through time. From the epic poetry, a simple oral tradition with a huge social component; to theatre, a performed art still embedded in community sense; and to novel, an individual and introspective form of narrative.

It is important to point out, historically speaking, writers constantly break standard and classical rules in art, specially since the 20^{th} century with the modernism. But besides that, novels are still bound to base elements needed to tell a story and these elements interact in a similar way.

In regards to novels, Candido [11] says a novel is always composed by three parts: the characters, the plot and the ideas. The "ideas" are the intentions and objectives the author brings to the story, by definition a very subjective concept, open to interpretation depending on the reader's own mindset. Due to this perspective, we removed this part from our base-questions, focusing on the two remaining parts, the character and the plot. It is important to point out that these two elements are also present in Aristotle's studies around drama, as mentioned before.

Therefore, our five base-questions were divided into 2 questions related to the characters within the book; 2 questions related to the plot those characters are inserted into; and 1 question related to the quantity of pages the book possesses.

2.1 Characters' questions

The evolution of drama to novels through the centuries brought a significant change to the characters' personalities. As theatrical characterization was no longer needed, the authors now had prospects to explore deep aspects of personality. So if before writers had around 2 or 3 hours to develop all characters and the plot, now they could develop as much as they wanted throughout the pages. This scenario and the classical differences between tragedy and comedy in theatre gave birth to two types of characters: flat and round character.

Candido [11] defines the flat characters as those with distinctive traits, strongly chosen in a way they can be quickly reminded by the reader when the author invokes a related scene. Furthermore, these traits do not change throughout the story. This type of character is quite associated with caricature and it normally gives a more simple and comical feeling.

For the round characters, Candido [11] describes them as more mysterious and deep, with more subtle personality traits which can evolve and change as the story goes and the character interacts either with other characters or the plot.

These terms can rapidly be associated with more simplistic ones: flat to simple characters (as the reader can comprehend them more easily) and round to complex characters (as it requires more analysis from the reader to understand their motivation). Hence, we chose the popular names, instead of the academic ones as they are easier to the public to connect as we wanted.

With this division, the first base-question is:

About the characters within the book, are they simple or complex?

Aristotle said very little about the virtue of the characters, considering it was not the most important trait in the story. For him, the main character's personality should be someone neither completely evil, nor completely good. But one thing he did not cease to say was that the tragedy should be build up around superior men, heroes.

The image of the Greek hero is a notable and popular concept (Aristotle mentioned plenty of them, including Orestes, Odysseus and Achilles). As a counter-image, the anti-hero was born in the Latin comedy, as a mocking version of the ancient heroic figure. In Pividori [12], the authoress briefly explains how the anti-hero figure evolved through time. Of course, the meaning of "hero" and "anti-hero" has changed a lot from Aristotle's time to ours, but these characters can still be identified and are present in the literature [13].

Therefore, we can define the second base-question as:

About the main character within the book, would they be considered a hero or an anti-hero?

It is important to point out that some modern novels have different characters with apparent same importance in the story. We decided not to indicate or classify what a main character should be, understanding that by letting this question open to readers, we might reach a common ground in the collective wisdom. Users than choose which is the main character within the book and answer it taking into account this persona.

2.2 Plot flow

Historically, the plot is significantly more relevant in drama than in novels and this is due to what we have mentioned before. In theater, playwrights have around two hours to perform their story, when in prose fiction, writers can go as far as they want to develop their characters. In fact, Aristotle claims that the plot is the most important aspect of a tragedy, considering that a tragedy is not an imitation of humans *per se*, but of actions. Happiness, sadness and even the ending of life is an action, not an aspect of life, where the characters are players of those actions.

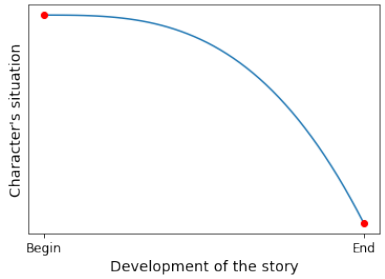
Hence, he analysed closely the flow of actions in the tragedies and asserted that a tragedy must always present an unbalanced flow of actions: going either from a good to bad situation, or from a bad to good situation.

However, even though we knew this topic had to be incorporated into the base-questions, adaptations would be needed to fulfill the aspects of novels for two reasons: first, modern novels might have more complex plots, with more characters and each one facing

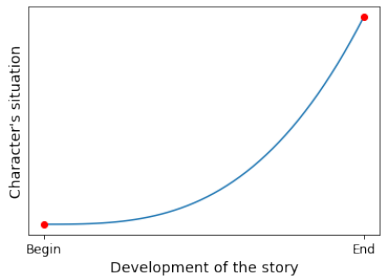
their own flow of actions (a topic we will address in the next base-question); and second, the concept of good and bad are more subtle nowadays.

In order to present this concept to the readers adequately, we choose to use two figures, each one related to a flow of actions, to set a common ground of understanding. Besides, we used the same main character resource we used for the second base-question. And the third base-question is:

About the main character's plot within the book, which figure better represents it: Figure 1a or Figure 1b?



(a)



(b)

Figure 1: Images of the third base-question

It is undeniable how refined and noble tragedies were, and are, considered. But by their own form, they are limited to their theatrical component. They cannot represent many parts of actions as they normally show only one flow of actions. Aristotle addressed this problem and said that a tragedy will not be, and should not be, able to fulfil all narrative aspects. But he presented a solution: the epic poetry.

Considering novels being a narrative form of literature, able to show different flows of actions, the fourth base-question is defined as:

About the narrative construction within the book, would it have several focus, that may or may not meet along the story, or an unique focus, without knowing with details what happens outside this center?

2.3 Quantity of pages

Pages is not a characteristic that critic authors have any concern or academic interest. It is possible to find plenty studies about *Les Misérables*, from Victor Hugo, with its 1, 500 pages (depending on the edition), and the same quantity about *The Death of Ivan Ilyich*, from Liev Tosltoy, with its 90 pages.

However, it is a very important attribute to the general public. It can be said that for most of the people, Victor Hugo's novel would be much less appealing than the Liev Tosltoy's one, simply because of the quantity of pages in each of them. Therefore, a question about this matter was included in the base-questions.

Following the same model of the other base-questions, it was important to keep the query binary. Furthermore, we did not want to influence the users by giving a direct division between the number of pages, understanding that each person has a particular standard to classify if a book has a lot or few pages. Thus, the fifth base-question was defined as:

In what concerns the length of the book: do you consider it has many or few pages?

3 DATA COLLECTION

As we wanted to use the base-questions not only as a search mechanism for recommendation, but also as a basis to generate our recommendations, we would need a database in a very specific way. For that, we have conducted a data collection with the public. In fact, the construction of datasets is been constantly explored by the scientific community, in Kuwaki et al. [14], the authors built a dataset to perform Sentiment Analysis oriented by supermarket reviews written in Portuguese.

The main goal of our algorithm is to group books to form a standard-evaluation based on users' opinions and apply this to recommend it to other users. Therefore, our database had to be focused on quantity of evaluations rather than on quantity of books.

To ensure we would gather data in this shape and also increase engagement with the public, we selected a base list. The fifty selected books were chosen mixing information from UK [15] and Brazil [16] (data information as it was in April, 2022). Besides this, some additional rules were defined to the list, to either increase engagement, or to provide diversity to it:

- An author must not have more than one book within the list. When facing this scenario, we chose the most popular novel;
- Book series or franchises were treated as a single book. Throughout the data collection, we instructed the public to give answers if they had read at least one book of the series, not requiring that they had read all of them.

The fifty books were divided in three balanced batches, so each batch would be as diverse as possible in terms of genre, consequently increasing user's connection to the list. The first batch contained ten books and the other two, twenty each.

The data collection was made in Google Forms. In each forms, the user was presented to a book and had to answer if they had ever read it or not. If they had, the form opened the base-questions to the user classify the novel and also a rating question based on their own opinion with a score from 1 to 5. If they had not read it, the form goes to the next book.

All the forms were vehiculated through the author's social media profiles and the links were shared as many time as possible until the deadline, September 1^{st}, when the forms were closed to not receive more answers. Table 1 shows the results obtained in the data collection with the public.

Batch	Start	People	Inputs	Average
1	April 25^{th}	91	221	2.43
2	May 8^{th}	48	246	4.50
3	June 20^{th}	39	107	2.74
	Total	113	544	4.81

Table 1: General engagement of the data collection

Although the people's engagement has dropped from batch to batch, the reader's inputs had its peak in the batch 2 and decreased significantly in the batch 3. The average inputs per person of each batch shows that the third one was better than the first, even though the number of people responding to the research had been smaller.

As mentioned before, we were looking for a high number of evaluations per book rather than a big number of books with few evaluations. Therefore, our expectancy was to have a distribution more skewed to the left, showing a good rate of inputs per book. Unfortunately, figure 2 shows that the data collection did not achieve this goal.

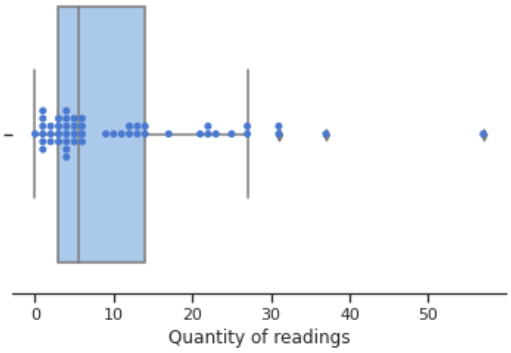


Figure 2: Readings distribution through the selected books

According to figure 2, 75% of the selected books had less than 14 inputs. Due to this poor distribution and imbalanced sample, strategies would have to be adopted to prevent dirty data from entering into the algorithm or the simple removal of these books from the recommendation. The two strategies we explored and that will be discussed in the next section are: oversampling through SMOTE [17], and through SELC, a novel method developed by the authors to deal with the sample.

4 PROPOSED ALGORITHM

The data collection provided us with a list of evaluations, but as our questions and forms were designed for a better user experience, all answers were texts which needed proper treatment.

As the base-questions were binary, we assigned 0 or 1 to each answer of each question. The mapping we used can be found in the table 2. Other data treatments and cleaning needed were done in Google Sheets.

With this, we can see the user evaluation as a map $\mathfrak{T}$,

$$\mathfrak{T} : \mathbf{U} \times \mathbf{B} \to \mathbf{B} \times \{0,1\} \times \{-1,0,1\}^5 \times \mathbb{5},$$

# Base-question	Answer	Binary digit
1	Simple	0
	Complex	1
2	Hero	0
	Anti-hero	1
3	Figure 1a	0
	Figure 1b	1
4	Several focus	0
	Unique focus	1
5	Many pages	0
	Few pages	1

Table 2: Answers mapping from text to binary

where $\mathbf{U}$ is the set of users who participated in the data collection (ie. $|\mathbf{U}| = 113$), $\mathbf{B}$ is the set of books (ie. $|\mathbf{B}| = 50$) and $\mathbb{5} = \{1, 2, 3, 4, 5\}$. This function $\mathfrak{T}$ maps $\mathbf{U} \times \mathbf{B}$ as follows:

$$(u,\ b) \mapsto \begin{cases} (b,\ 0,\ (-1)^5,\ n) & \text{if user } u \text{ had } \textbf{not} \text{ read the book } b \\ (b,\ 1,\ v,\ n) & \text{if user } u \text{ had read the book } b \end{cases}$$

where v is a binary 5-vector and $n \in \mathbb{5}$ related to the rating of the book by the user.

As our goal was to recommend, at least, five books based in the user's input, we can write the recommendation task $\mathfrak{R}$ as

$$\mathfrak{R} : \mathbf{U} \times \{0,1\}^5 \to \mathcal{P}^5_{\mathbf{B}},$$

where $\mathcal{P}^5_{\mathbf{B}}$ is a subset of the power set of $\mathbf{B}$ where the elements have cardinality equal 5, ie. $\mathcal{P}^5_{\mathbf{B}} = \{X \subset \mathbf{B} \mid |X| = 5\}$.

With this, we can now discuss each step (marked with a red number) of the algorithm as shown in the Figure 3.

1. Cluster centres. The centres are the first part of the algorithm and they are the basis from where all development follows.

The centres are calculated per novel, so each novel has its own centre which changes as more information is collected from users on the book. They are obtained from the weighted average of the data, with the base-questions answers as main source and the user's evaluation as the weight. We can define the centres map as

$$\mathfrak{C} : \mathbf{B} \times \{1\} \times \{0,1\}^5 \times \mathbb{5} \to \mathbf{B} \times \mathbb{R}^5,$$

such that

$$\mathfrak{C}(b,\ 1,\ v,\ k) = (b,\ \overline{v}),$$

where

$$\overline{v} = \frac{1}{\sum k} \sum k.v,$$

for all v and k associated with the novel b.

From the notation we are adopting, the map $\mathfrak{C}$ follows directly from $\mathfrak{T}$ image, in a way that we only use the data from users who have read the novel (therefore the $\{1\}$ in the second entry in the domain).

2. Context pre-filtering. Now that we have the set of novel centres, $\mathbf{C} = \text{Img}(\mathfrak{C})$, the algorithm base is completed. With that, the user can find books that match theirs desires, so basically comparing their input, $\mathcal{I} \in \{0,1\}^5$, with all centres and recommending novels that are close to this input.

The novel match is done using a Euclidian distance function between $\mathcal{I}$ and all the book within $\mathbf{C}$. The smaller the distance, the more accurate the book is with the user's profile. An important

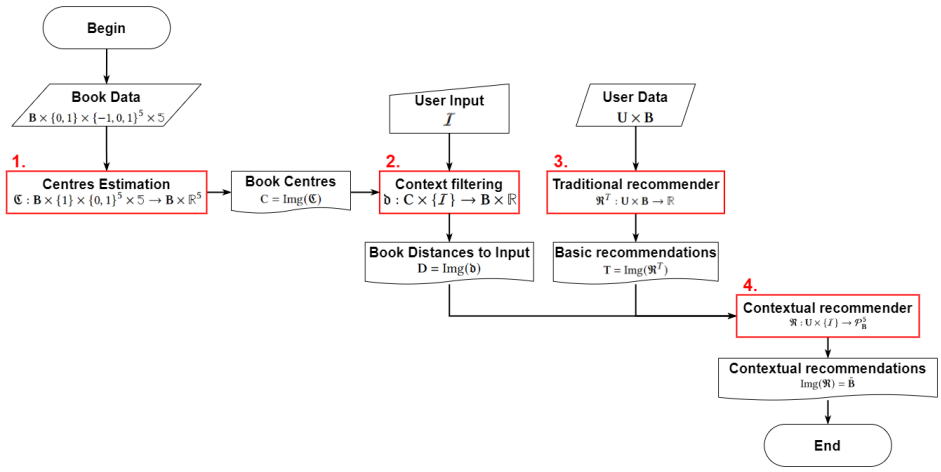


Figure 3: Description of the proposed algorithm

observation must be made about the way we interpret $\mathcal{I}$, more specifically the last base-question within it.

Regarding the question about the number of pages, readers had to select if the novel had many or few pages. But when searching for a book, although the user still had to indicate if they wanted a book with many or few pages, we decided to interpret the "many pages" as "it does not matter". This decision was done taking into account that people might look for a small book due to personal preferences, but no one proactively wants a big book, instead it makes no difference to them.

Therefore, the distance function $\mathfrak{d}$ is such that

$$\mathfrak{d} : \mathbf{C} \times \{\mathcal{I}\} \to \mathbf{B} \times \mathbb{R},$$

where

$$(b,\ v,\ \mathcal{I}) \mapsto \begin{cases} \left(b, \sqrt{\sum_{i=1}^{5} (v_i - w_i)^2}\right) & \text{if } w_5 = 0 \\ \left(b, \sqrt{\sum_{i=1}^{4} (v_i - w_i)^2}\right) & \text{if } w_5 = 1 \end{cases}$$

with $\mathcal{I} = (w_1, \ldots, w_5)$.

The list of distances provided by $\mathfrak{d}$ would be enough to continue the development. However, as we mentioned earlier, a problem arrives from the fact that two novels, b_1 and b_2, might have distances from $\mathcal{I}$, ρ_1 and ρ_2, such that $\rho_1 < \rho_2$, but $|\mathfrak{T}^{-1}(b_1)| \ll |\mathfrak{T}^{-1}(b_2)|$, which means that b_1 received significantly less evaluations than b_2.

This issue comes from the fact that our dataset has an imbalanced profile in the perspective of the evaluations per novel. To deal with this and provide better filter to our data and more accurate recommendations, we applied two methods: SMOTE (acronym for *synthetic minority oversampling technique*) and SELC (acronym for *strict evaluation list control*), a novel proposition by the authors.

While SELC technique was designed by the authors as an *ad hoc* approach and easier to implement model, SMOTE is a popular technique that uses KNN (or other clustering algorithms) to create more samples of an imbalanced classes (initially developed in Chawla et al. [18], but for a more complete review, see Fernández et al. [17]).

The SELC model was created to filter the data with few evaluations and only include in the following parts of the algorithm those that are extremely accurate in terms of distance from $\mathcal{I}$. The implementation can be seen in Figure 4. It works by taking the median of the number of evaluations,

$$M = \text{med}\ \{|\mathfrak{T}^{-1}(\{b\} \times \{1\} \times \{0,1\}^5 \times \mathbb{5})| \mid \forall b \in \mathbf{B}\},$$

and using it to split into the set of novel in two lists:

$$\mathbf{B}^h = \{b \in \mathbf{B} \mid |\mathfrak{T}^{-1}(b)| \geq M\} \quad | \quad \mathbf{B}^l = \{b \in \mathbf{B} \mid |\mathfrak{T}^{-1}(b)| < M\},$$

so $\mathbf{B}^h$ has all books with more than M evaluations and $\mathbf{B}^l$ has all books with less than M evaluations.

With the user input, $\mathcal{I}$, we calculate $\mathfrak{d}$ in each set. We define $m = \min \mathfrak{d}(\mathfrak{C}(\mathbf{B}^h) \times \{\mathcal{I}\})$, the smallest distance and the minimum distance a book from $\mathbf{B}^l$ must have to be included in the final list. Therefore,

$$\mathbf{B}^r = \{b \in \mathbf{B}^l \mid \mathfrak{d}(b) \leq m\},$$

and the filtered list is $\mathbf{B}^f = \mathbf{B}^h \cup \mathbf{B}^r$.

The SMOTE method is far more complicated to implement in our scenario. It is important to point out that the technique is not applied after the centres calculation and part of the data filtering, but instead it is applied in the first part of algorithm. As it consists of a resample technique, SMOTE is used to increase the amount of data in the raw set $\mathbf{B} \times \{1\} \times \{0,1\}^5 \times \mathbb{5}$. The SMOTE implementation diagram can be seen in the Figure 5.

Let be $M = \max |\mathfrak{T}^{-1}(\mathbf{B} \times \{1\} \times \{0,1\}^5 \times \mathbb{5})|$, the maximum of evaluations per book. This value will be used as the standard to resample the other novels so they can all have the same number of evaluations.

As SMOTE resamples data per class we need to organize our basis to resize each one. To perform this, we will use the ratings provided by the users when evaluating the books. For any $b \in \mathbf{B}$, let be $e_i = |\mathfrak{T}^{-1}(\mathbf{B} \times \{1\} \times \{0,1\}^5 \times \{i\})|$, for $i \in \mathbb{5}$, the number of evaluations of the novel b for the class i. Hence,

$$e = e_1 + e_2 + e_3 + e_4 + e_5$$

is the total number of evaluations.

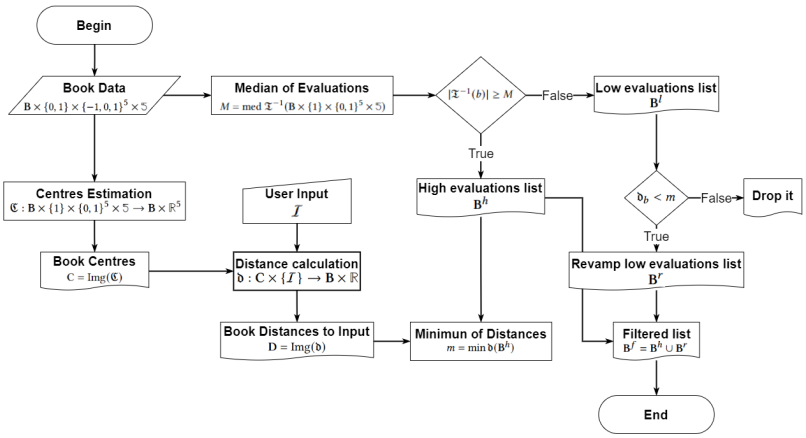


Figure 4: Description of the SELC model

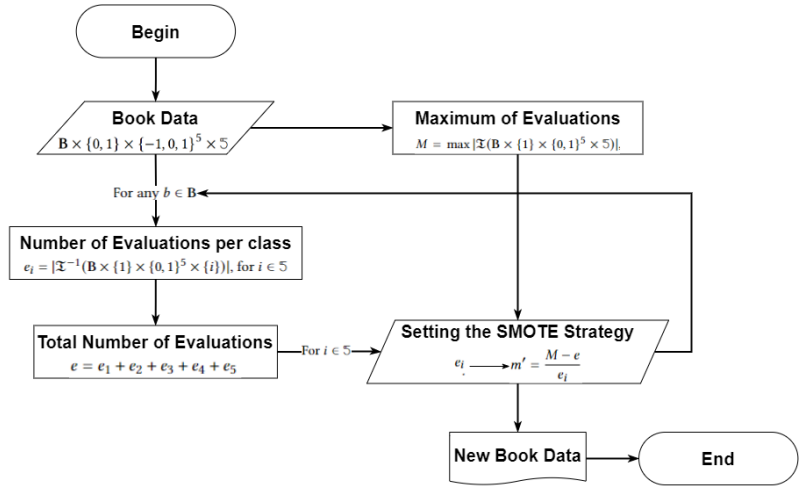


Figure 5: Description of the SMOTE model

As we want all novels to have the same amount of samples as the biggest one, ie. M, the data to input for $b \in \mathbf{B}$ is $m = M - e$. The SMOTE strategy to complete the task per class is

$$m' = \frac{m}{e_i},$$

for any $e_i \neq 0$ and $i \in \mathbb{5}$.

With this, we can apply $\mathfrak{C}$ to obtain the centres in a basis where all novels have the same amount of evaluations.

3. Traditional recommendations. From this point, independent of the implementation we chose, SELC or SMOTE, we have a set of books and its distance from $\mathcal{I}$. To improve the recommendations, we decided to apply a traditional content-based recommendation algorithm, $\mathfrak{R}^T$.

The algorithm we chose is the one from the Surprise package [19]. The Surprise algorithms provides us predictions of user's ratings of novels they have not read and these values will be used in the last part of our model.

Using the built-in function GridSearchCV (with 5 k-folds) within Surprise, we have searched for the algorithm and parameters which better fit our data and have the best RMSE result.

The selected combination was from the KNNBaseline algorithm and the following parameters

Parameter	Value
Similarity measure	Cossine distance
Similarity level	Items only
Maximum neighbours	30
Minimum neighbours	3

The Surprise algorithm was implemented as

$$\mathfrak{R}^T : \mathbf{U} \times \mathbf{B} \to \mathbb{R}$$

and now we can combine the ratings prediction and input distances to generate the recommendations.

4. Contextual recommendations. The final step of the algorithm is the combination of the distances between the novels centres and the user input $\mathcal{I}$, Img($\mathfrak{d}$), with the tradition recommendation for

the user, $\mathfrak{R}^T$. The contextual recommendation task is

$$\mathfrak{R} : \mathbf{U} \times \{\mathcal{I}\} \to \mathcal{P}^5_{\mathbf{B}}.$$

So let be $u \in \mathbf{U}$, an user searching for novel recommendations based on their input $\mathcal{I}$. From the input, we have

$$\mathfrak{d} : (b,\ v,\ \mathcal{I}) \mapsto (b,\ \rho)$$

the distances. And from the user, we have

$$\mathfrak{R}^T\Big|_u : (u,\ b) \mapsto \kappa$$

the traditional recommendations. Due to the constructions made, we have that $\rho \in [0, 1]$ and $\kappa \in [1, 5]$. With this we can create a score by novel based on these numbers

$$\sigma = \frac{\kappa}{\rho}.$$

The score σ provides a perfect balance between the distances and ratings, considering that $\lim_{\rho \to 0} \sigma = \infty$, as the closer the novel is from $\mathcal{I}$, the bigger the score.

Therefore, we can order $\mathbf{B}$ using σ. Let be $\mathbf{B}_\sigma$ with

$$\mathbf{B}_\sigma = \{b \in \mathbf{B} \mid \sigma_1 \geq \cdots \geq \sigma_n\},$$

where σ_i is the score of the novel b_i. And we can pick the first five novels in $\mathbf{B}_\sigma$ to form our recommendations, $\tilde{\mathbf{B}} = \{b_1, b_2, b_3, b_4, b_5\} \in \mathcal{P}^5_{\mathbf{B}}$.

The set $\tilde{\mathbf{B}}$ is such that its elements are novels that have good proximity to the user's desire as well as a high probability the user will give a good rating. The recommendation list is

$$\mathrm{Img}(\mathfrak{R}) = \tilde{\mathbf{B}}.$$

5 MODEL EVALUATION

A proper evaluation of the algorithm would require a qualitative research with the participants of the data collection, so they could rate the recommendations on both accuracy with their wishes and personal satisfaction. But we can use the very data collection to determine in some levels how the recommender performs.

To evaluate the algorithm, we will transform the recommender in a binary classifier. Hence we use the user's rating to create two classes: if the rate is between 1 and 3, inclusive, the novel is considered "not recommendable for the user" (and is labeled as 0) ; if the rate is between 4 and 5, the novel is considered "recommendable for the user " (and is labeled as 1). Therefore, we have a set like

$$\mathbf{B} \times \{1\} \times \{0, 1\}^5 \times \{0, 1\}.$$

With this mechanism, we can evaluate the algorithm using binary classification scores.

The main idea of any recommendation is to name contents the user had not seen yet. Therefore, the suitable novel recommendation to user $u \in \mathbf{U}$ would be:

$$\tilde{\mathbf{B}} = \mathbf{B}_\sigma \cap \mathfrak{T}\Big|_u^{-1}(\mathbf{B} \times \{0\} \times \{-1\}^5 \times \mathbb{5}).$$

Which means we would only recommend novels the user has never read. But to evaluate the model, we will consider other type of recommendation: **mixed novels**, where the algorithm is able to recommend any novel, read it or not. This is made so we can have a closer experience to a qualitative research.

Using the data from the data collection, we applied the algorithm for each user and each vector of base-question they provided. If the book is in the list of five recommendations, we label 1; otherwise, we label 0. This operation can be seen as a map

$$\mathfrak{A} : \mathbf{U} \times \mathbf{B} \to \{0, 1\}$$

and

$$\mathfrak{A} : (u, b) \mapsto \begin{cases} 1 & \text{if } b \in \mathfrak{R}(u, v) \\ 0 & \text{if } b \notin \mathfrak{R}(u, v) \end{cases}$$

where $v \in \{0, 1\}^5$ is such that $\mathfrak{T}(u, b) = (b, 1, v, k)$, with $k \in \mathbb{5}$. Basically, we are using the own evaluation provided by the user as a context input into the algorithm.

The procedure allows us to compare this output and the recommendable classification made earlier. In short, we want to be able to recommend the novels the user liked it and do not recommend the ones they did not. Moving forward, we can compare the two approaches we had proposed before: SELC and SMOTE.

The Cohen's Kappa shows that the lists from $\mathfrak{A}$ using SELC and SMOTE have significantly agreement, $\kappa = 0.72$. Hence, both methods generate similar lists.

The measure we use to determine the accuracy of each model is the Matthews Correlation Coefficient (MCC). As Chicco et al. [20] state, the MCC provides a more truthful and informative score for binary classifiers than the Cohen's Kappa and Brier Score. Comparing each method predictions from $\mathfrak{A}$ with the true values we have the following results.

	SELC	**SMOTE**
MCC	0.175	0.154

For MCC, SELC has better results than SMOTE, around 13% bigger. But considering the range of MCC is from −1 to 1, both of them have poor results. However, as we are handling a binary classifier, there are some measures we can use to better understand.

The confusion matrix of the methods might provide some insights.

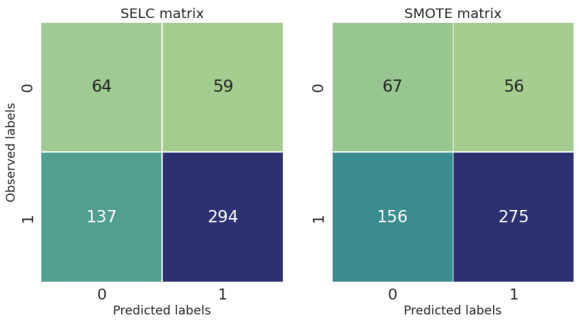


Figure 6: Confusion matrix of the models

Figure 6 shows us that the main problem for both algorithm is in the *false negative* labels, 69% of errors comes from them for SELC and 73% for SMOTE. Even so, they still have good *weighted recall*.

	SELC	**SMOTE**
MCC	0.17	0.15
Recall	0.65	0.62

For the purpose of the analysis, to have a high false negative rate is not necessarily bad. As we are recommending both read and unread novels, the fact that we are not recommending books they liked means that we are recommending some they might like even more. Most important is to assure we are not recommending novels they did not like. On this matter, both algorithms are doing well, as we can see on the weighted precision and F1-score.

	SELC	SMOTE
MCC	0.17	0.15
Recall	0.65	0.62
Precision	0.72	0.71
F1-score	0.67	0.65

With this we can conclude that although the MCC scores for both models are low, both of them are naturally consistent to what we want from a recommender. Moreover, the other measures point to relevant algorithms with a clear advantage to the SELC method in opposite to the SMOTE one.

6 CONCLUSIONS AND FUTURE WORK

Independently of the approach, the algorithm to recommend novels is properly defined and works its purpose to bring recommendations according to the user's desire.

The structure allows to overcome the New User problem, since the pre-filtering process would be able to provide recommendations even for a user with no novels evaluations. Furthermore, both SELC and SMOTE methods were designed to deal with the New Item problem. And the active search makes impossible to the algorithm to produce serendipitous recommendations or reach overspecialization.

A future development would be making a qualitative research. By reaching some of the users who have participated in the data collection, providing them recommendations based on their desires, and let them evaluate the quality of the recommended list, seeking to understand if it is accurate to what they were looking for and if the novels seem enjoyable to them. As SELC has showed a better performance, we would focus only in list supplied by this method.

Another relevant discussion is how much scalable the proposed algorithm is to other media.

In what concerns to cinema, the application would be not only feasible, but also simple. Cinema is a modern evolution to theatre, another form of art with the object to tell a story. Similar to novels, films and plays are build up with the same elements: characters that interact around a plot, which affects these characters and changes their personality. As we mentioned before, Aristotle [8] is focused on the theatrical scenario and besides the other two previous topics, we would have to add something about spectacle (most associated with cinematography) and diction (referring to stage and acting performance). Therefore, by adding some two or three binary base-questions it would be possible to use the same framework and generate film recommendations.

Regarding music, the challenging would be much bigger. The proposed algorithm is based on the binary questions and while those questions can provide some basic information about their specifics, they do not account for the complexity and nuances of music. Music encompasses a variety of factors, such as genre, tempo, lyrics, and mood. Moreover, it does not have the storytelling structure novels and cinema have, and we largely used to build the algorithm.

ACKNOWLEDGMENTS

This work was parcially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação Amazônica de Amparo a Estudos e Pesquisas (FAPESPA) - PRONEM-FAPESPA/CNPq nº 045/2021. We would like to acknowledge Adriane Delgado, for the literature and critic studies references; and the reviewers, for the suggestions which improve our paper.

REFERENCES

[1] Umberto Panniello, Alexander Tuzhilin, and Michele Gorgoglione. Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction*, 24(1):35–65, 2014.
[2] Negar Hariri, Bamshad Mobasher, and Robin Burke. Context-aware music recommendation based on latenttopic sequential patterns. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 131–138, 2012.
[3] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.
[4] Zeinab Abbassi, Sihem Amer-Yahia, Laks VS Lakshmanan, Sergei Vassilvitskii, and Cong Yu. Getting recommender systems to think outside the box. In *Proceedings of the third ACM conference on Recommender systems*, pages 285–288, 2009.
[5] Umair Javed, Kamran Shaukat, Ibrahim A Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, 2021.
[6] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.
[7] Rudson Francicso da Silva Mendes and João Victor Ribeiro de Jesus. Exploração de modelos de aprendizado de máquina e seleção de atributos para employee attrition. *Anais do Computer on the Beach*, 12:267–272, 2021.
[8] Aristotle. *Poética.* Editora 34, 2015.
[9] Ian Watt. *The rise of the novel.* University of California Press, 2001.
[10] Marcela F.C. Oliveira. *Em busca do sentido perdido: expressões literárias da queda da experiência moderna no pensamento de Walter Benjamin.* PhD thesis, Pontificia Universidade Católica do Rio, 2009.
[11] Antonio Candido. A personagem do romance. In *A personagem de ficção*, pages 51–80. Perspectiva, 2000.
[12] Cristina Pividori. *The Death and Birth of a Hero: The Search for Heroism in British World War One Literature.* PhD thesis, Universitat Autònoma de Barcelona, 2012.
[13] Rosette C Lamont. From hero to anti-hero. *Studies in the literary imagination*, 9 (1):1, 1976.
[14] Vinícius Takeo Friedrich Kuwaki, Mateus Nepomuceno Ladeira, Matias Giuliano Gutierrez Benitez, and Rui Jorge Tramontin Junior. Building a corpus from supermarket reviews in portuguese for document-level sentiment analysis. *Anais do Computer on the Beach*, 13:119–125, 2022.
[15] The Guardian UK. The top 100 bestselling books of all time: how does fifty shades of grey compare? Last access in August 14^{th}, 2022, 2012. URL https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-compare#data. Unable to identify the author.
[16] Amazon Brazil. Mais vendidos em clássicos de ficção, 2022. URL https://www.amazon.com.br/gp/bestsellers/books/7872689011/ref=zg_bs_nav_books_2_7872706011. There is no author.
[17] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.
[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
[19] Nicolas Hug. Surprise: A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174, 2020. doi: 10.21105/joss.02174. URL https://doi.org/10.21105/joss.02174.
[20] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, 9:78368–78381, 2021.