Models of Literary Evaluation and Web 2.0. An Annotation Experiment with Goodreads Reviews

Simone Rebora[1,*,†], Gabriele Vezzani[1,2,†]

[1]University of Verona
[2]RWTH Aachen University

Abstract

In the context of the Web 2.0, user-genrated reviews are becoming more and more prominent. The particular case of book reviews, often shared through digital social reading platforms such as Goodreads or Wattpad, is of particular interest, in that it offers scholars data regarding literary reception of unprecedented size and diversity. In this paper, we test whether the evaluative criteria employed in Goodreads reviews can be included in the framework of traditional literary criticism, by combining literary theory and computational methods. Our model, based on the work of von Heydebrand and Winko, is first tested through the practice of heuristic annotation. The generated dataset is then used to train a Tranformer-based classifier. Last, we compare the performance of the latter with that obtained by instructing a Large Language Model, namely GPT-4.

Keywords

literary evaluation, digital social reading, annotation, transformer models, LLMs

1. Introduction

Nowadays, reviews are ubiquitous. Pushed by our natural tendency to share information and encouraged by the very companies that sell us their products, we constantly take part in the production and accumulation of huge amounts of data regarding our preferences and judgements. Literature has also been strongly affected by this phenomenon. On digital social reading platforms [22, 19], such as Goodreads, one can find terabytes of information regarding the reception of millions of books, covering the tastes of the most diverse typologies of readers. Finding efficient ways to manage these records is essential not only for market reasons – as demonstrated by the great effort that companies put into developing algorithms to better capture and predict users' tastes[1] – but also for research purposes. Never before, in fact, has it been possible to gather insights about the reception of literary works as extensive and diverse

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
simone.rebora@univr.it (S. Rebora); gabriele.vezzani@univr.it (G. Vezzani)
0000-0002-1501-3774 (S. Rebora)

[1]Although the specificities of the recommendation algorithms implemented by Amazon are not shared with the general public, at the following link it is possible to access an updated list of the algorithms (with relative publications) that obtained the best performance over the Amazon-Book dataset (a dataset containing information for over 3 million user generated book reviews): https://paperswithcode.com/sota/recommendation-systems-on-amazon-book

as the ones that can be drawn from these platforms. Designing ways to analyze this data will allow us to shed a whole new light on the dynamics that regulate literary evaluation.

Literary studies have been preoccupied with evaluation since around the 1980s, when, thanks to the influence of the emergent field of postcolonial studies [2, 25], it started to become more and more evident that the prominence of the works of the so-called western canon wasn't at all objective, but rather an arbitrary construct, contingent on the social and cultural background of the readers. Van Rees [28] proposed a model in which literary value was built through a three-step process involving journalistic reviewers, essayists, and academic critics. At the same time more detailed and flexible, von Heydebrand and Winko's model [13] also posits the social construction of literary value, although allowing for the contribution of any agent within the literary field. Similar ideas are shared by Dalen-Oskam [6], who recently demonstrated, in an empirical fashion, how social biases can influence the evaluation of books.

According to these authors, society influences literary evaluation by setting up standards against which books are judged. This would imply the existence of criteria independent from any specific evaluative act, that manifest themselves across a multiplicity of such acts, thus offering a means to systematize them. In fact, as von Hydebrand and Winko have shown, although they can vary greatly depending on the historical context, these "standards of value" can nonetheless be organized according to some general features: they can concern the aesthetic level of a book (its style, characters, plot, etc.), the way it relates to other works (thus being original or derivative, for instance), the impact it has on its reader, and so on.

In this paper, we build on von Heydebrand and Winko's theory and interpret online reviews as acts of "linguistic evaluation". What distinguishes such acts from more implicit forms of evaluation (even the simple purchase of a book could be interpreted as such) is that they "require a standard of value - as well as certain categorizing assumptions - in order to progress from the description of a text to its evaluation" [12, p. 227]. Seen as evaluative acts of this kind, book reviews are not simple expressions of individual taste, but the result of a socially learned practice with specific schemes and regularities. Different kinds of evaluative criteria can then be taken as the axis along which to situate single reviews, thus quantifying their variability and reducing their complexity. Furthermore, knowing which criteria are implemented by an individual (or a group) could allow us to predict their future judgements or to reveal the societal influences operating on them.

To make the identification of evaluative criteria as objective as possible, we turned to the practice of "interpretative markup", a form of annotation "devoted to recording a scholar's or analyst's observations and conjectures in an open-ended way" [20, p. 202]. As argued by Gius and Jake [10], this kind of annotation, when carried out in a collaborative way, can be a powerful tool for systematizing the interpretation of potentially ambiguous texts. Our main goal consisted in using the annotated dataset for training a classifier capable of recognizing the use of different evaluative criteria in possibly any online book review.

The paper is organized as follows. In section 2, we present the corpus that constituted the basis of our work. The discussion of the annotation process occupies Section 3, where we outline the tagset we developed (3.1), the workflow we followed and the main problems we encountered (3.2). To automate the annotation process we tried two different approaches: fine-tuning a Bert-based classifier and instructing a Large Language Model (GPT 4). The results of these approaches are presented, respectively, in Sections 4.1 and 4.2.

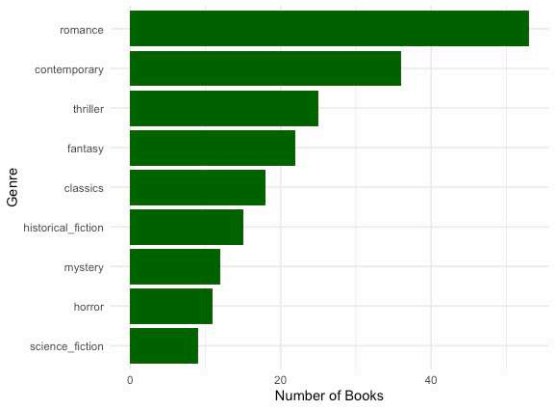

Figure 1: Distributions of books per genre

2. Data

The dataset used to perform the annotation task was built on the same basis of the AbsORB dataset: a selection of "approximately six million English language reviews of nine different genres (i.e., fantasy, romance, thriller, horror, mystery, science fiction, historical fiction, contemporary, classics)" [15], downloaded from the Goodreads website between 2018 and 2019. Out of them, a total of 100 reviews were randomly selected, with just one filter set to exclude extremely short (below 200 words) ones. To counter an intrinsic imbalance in the dataset (i.e., the dominance of reviews of "young adult" novels, frequently tagged as "fantasy" or "romance"), we added a filter which forced the selection of 20 reviews from among the ones tagged as "classics". Reviews were then automatically split into sentences by using SpaCy[2], so as to allow a sentence-by-sentence annotation.

A first partitioning of our corpus, composed by 11 reviews, was used as a toy dataset to train the annotators, and was not therefore retained for analysis. The remaining 89 reviews, which constituted our main corpus, had a mean length of 1155 tokens (SD = 449). They were written by 84 different reviewers in a timespan going from 2008 to 2018. The number of reviews per year is lightly skewed towards more recent dates reflecting the growth of the Goodreads platform. There are 84 different books reviewed in the corpus, spanning across 9 different genres (with most of the books belonging to more than one genre). The distribution of reviews per genre and per year can be seen, respectively, in Figures 2 and 1.

All copyright and privacy implications in using such a dataset have already been discussed [16, 23]. In any case, to safeguard the rights of the authors of the reviews and to comply to copyright limitations (while still profiting from the research exceptions recently introduced in multiple European legislations, following the 2019 *Directive on Copyright in the Digital Single Market*), we decided not to share it publicly. Researchers who would like to access it will have to contact the authors of this paper, by stating their intended use.

[2]https://spacy.io/

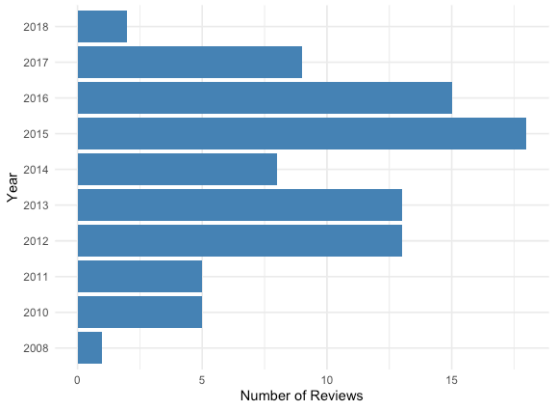

Figure 2: Distribution of reviews per year

3. Annotation

3.1. Designing the Tag Set

The starting point for the creation of our tag set was von Heydebrand and Winko's model, which systematizes literary evaluation based on the different kinds of criteria that can be employed for such a task. However, in order to be adapted to our material and goals, it needed to be simplified. First of all, as demonstrated by the high number of categories they devised, it is clear that von Heydebrand and Winko aim at maximizing the exhaustivity of their model. Our intent was slightly different. In fact, we needed to account for the highest possible number of scenarios with the fewest possible categories, in order to reach a number of examples per category high enough to train a classifier, without having to annotate an extreme number of reviews.

Furthermore, despite their intent to take into consideration the contributions of potentially any actor within the literary field, it is quite clear that the kind of evaluative acts that von Heydebrand and Winko had in mind while designing their model were quite different from what one can find today on digital social reading platforms (and quite understandably so, given that their essay predates the advent of the latter by almost ten years). Many of their categories reflect the judgment patterns of, if not professionals, at least cultivated readers, capable of assessing the salient stylistic features of a given book, as well as its positioning in the overall landscape of literary tradition. On the other hand, online reviews "are mostly expressions of consumer satisfaction or dissatisfaction" [26, p. 3]. In them, the individual dimension is magnified, with evaluations hinging around the reader's personal relation with the book.

In our tag set, we devoted a special attention to criteria based on one's own thoughts, feelings and personal experiences. However, not all can be traced back to the individual level. By contrast, we believe literary evaluation to be a complex phenomenon, involving not only a reader and a book, but also the whole community in which the interaction between the former two takes place. To account for such complexity, we have incorporated in our tag set two further dimensions alongside that of individual criteria: one for evaluations based on the book

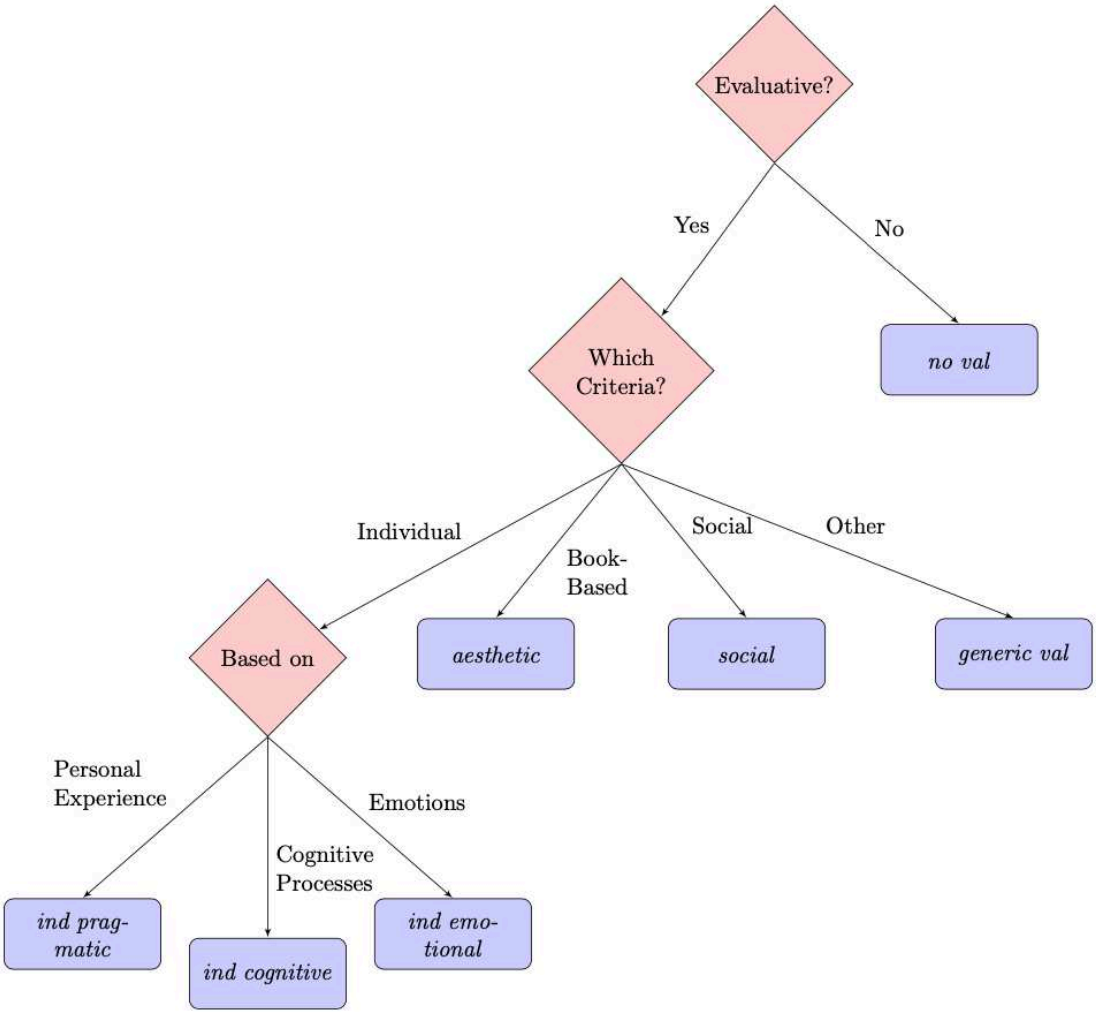

Figure 3: Chart of the annotation workflow

itself, its content-related or formal aspects, and one for considerations on its societal value, its impact on a given community of readers. Furthermore, we have included a label for evaluations that do not fall into any of the aforementioned categories. In Appendix A, we list the 7 labels of our tag set, with a short description for each. Figure 3, on the other hand, shows the overall structure of the tag set.

3.2. The Annotation Workflow

After a brief training regarding the task they were going to perform, two annotators were asked to independently annotate all the reviews in our corpus, attributing one (and only one) label to each of their sentences.

Table 1
Kappa values for different labels throughout the rounds

Label	Second round	Third round	Fourth round	Fifth round
all_labels	0.49	0.46	0.51	0.57
no_val	0.66	0.57	0.64	0.64
aesthetic	0.43	0.52	0.49	0.52
ind_cognitive	0.25	0.18	0.16	0
ind_emotional	0.39	0.05	0.27	0.57
ind_pragmatic	0.25	0.17	0	0
social	0	0	0	0.33
generic_val	0.35	0.32	0.44	0.47

The choice to take sentences as the basic unit of our analysis allowed us to organize the annotation in a very straightforward way, by providing annotators with tables where each row corresponded to one sentence, to which they had to associate a label in a dedicated column (columns containing the title of the reviewed book, the review and the sentence id numbers were also included). In the cases where an evaluation spanned across multiple sentences, annotators simply had to attribute the same label to each one of them.

A first partitioning of our corpus, composed by 11 reviews, was used as a toy dataset to train the annotators (annotations were therefore not retained for analysis). The actual annotation work was carried out on the remaining 89 reviews in 4 consecutive rounds, each followed by a meeting during which cases of disagreement were discussed, as well as possible shortcomings of our tagging system. Inter-annotator agreement was computed using Cohen's Kappa. The scores for each round are reported in Table 1.

Considering the entirety of the tag set ('all_labels'), the coefficients are in the range of a moderate agreement [18], while their increase over time can be taken as a sign of the efficacy of our workflow. Furthermore, the growth of the agreement for the tags 'generic_val' and 'aesthetic' can be seen as a consequence of the consecutive clarifications that we were able to offer during the meetings that followed each round. The agreement for the category 'no_val' (i.e., the simple distinction between evaluative and non-evaluative sentences), despite a drop in round 2, remained almost unchanged throughout the entirety of the work. It should be noted that this was by far the most frequent label, accounting for 26% of all annotations. This is why we managed to reach such a substantial coefficient for this category, which still registered many instances of disagreement, resulting for the most part from intrinsic ambiguities in our data. Take, for instance, even a sentence as simple as the following: "[this book] is not a horror novel"[3]. In and of itself, this would look like a mere observation. However, with a small interpretative leap, we could see it as a negative evaluation, a way for the reader to express their disappointment in finding out that the book was not what they expected it to be. To make the identification of evaluative sentences as unambiguous as possible, we then decided that for a sentence to be considered as such it needed to contain an explicit evaluation of the reviewed book, or an explicit mention of the impact it had on the reviewer. Although focusing only on explicit cases could be seen as limiting, such approach allowed us: a) to reduce disagreement

[3]https://www.goodreads.com/review/show/1373835925

between annotators, and b) to gather a corpus of clearly evaluative sentences to successfully train a classifier, if not in recognizing all of our categories, at least in distinguishing evaluative statements from other elements that can be found in a review, like summaries of the book's plot, accounts of reviewer's personal experiences, and so on.

The absence of agreement registered for the 'social' label in the first three rounds can be explained by the extremely limited number of occurrences of this category, which made it difficult to develop specific guidelines. To understand such scarcity, let's remember that the label was thought to account for references of socially established evaluative standards, such as prestigious prizes and canonical works. In many cases, such standards are the manifestation of an academic – as in the case of canonical works, see [11] –, or professional perspective on literature. They are the standards held by that cohesive community of readers that Chervel calls the "literary establishment" [5]. By not referencing the standards of the latter, Goodreads users simply demonstrate to belong to a different community, characterized by its own standards and rituals (think about the Goodreads Choice Award, a literary prize where winners are voted exclusively by members of the Goodreads community).

More than giving rise to a stable and universal canon, on Goodreads the social component of literary evaluation seems to operate at the level of specific literary genres. Indeed, both annotators reported noticing that reviews tended to follow genre-specific evaluative patterns. One of these patterns, concerning the over-represented genre of young adult novels (in Figure 2 split between romance and fantasy) could be at the heart of the problems we encountered with the labels in the 'individual' category, reflected in the erratic trends of the corresponding kappa coefficients. Something that became apparent already from the first rounds of annotation, is that readers of this genre tend to develop very strong personal relationships with the characters in the books they read, either treating them as if they were real persons or completely identifying with them. This phenomenon gave rise to an ambiguity that caused a good part of the disagreement between annotators. Take, for instance, the sentence "there is a depth to his character that bit by bit when revealed you can't help but fall that much harder for him because his character demands nothing less"[4]. Apart from the poor grammar, one notes right away that the reviewer is evaluating the way a character is built, which would make the sentence fall under the category 'aesthetic'. However, far from being a neutral assessment of a stylistic feature, the sentence is highly emotionally charged, which would justify the label 'ind_emotional'. Lastly, based on the patent involvement of the reader in what they are reading, one could argue for the attribution of this sentence to the 'ind_pragmatic' category. To face such ambiguity, we decided to tag as 'aesthetic' all the sentences that contained an explicit reference to literary art (e.g., plot, characters, style).

The last step after annotation has been a curation phase, during which we 'settled' the cases of disagreement between the annotators and attributed a definitive label to the respective sentences. This was necessary to allow for the subsequent phase, the training of a classifier, for which each sentence in our dataset needed to be assigned to one and only one label. In extremely rare cases (less than one percent of the sentences) we also intervened to correct attributions that, despite the annotators' agreement, were blatantly wrong.[5]

[4] https://www.goodreads.com/review/show/225468198

[5] The reason for the presence of such cases is not hard to find, given that, as we have said, several problems with

Table 2
Mean efficiency scores for the three models. A full report on the 5-fold cross validation can be found in our GitHub repository

Model	Metric	3-class	Binary
google-bert	Accuracy	**0.820**	**0.860**
	F1-macro	0.553	**0.832**
	F1-weighted	0.795	**0.860**
LiYuan	Accuracy	0.809	0.848
	F1-macro	0.528	0.818
	F1-weighted	0.782	0.850
JoelVIU	Accuracy	0.819	0.858
	F1-macro	**0.558**	0.826
	F1-weighted	**0.796**	0.858

4. Traininig the Classifier. BERT vs. GPT

The following step in our project relied on the annotated dataset to train a classifier able to recognize evaluative acts in large amounts of text. In recent years, two main approaches have become dominant in machine learning for text classification. The first implies fine-tuning base (or task-specific) Transformer models thanks to the knowledge stored in annotated datasets [14]. The second benefits from the flexibility of Large Language Models, instructing them to perform ex novo the annotation task based on a set of instructions (or examples) [21]. The materials created in our project offered a valid groundwork for both approaches.

Analyses have been carried out with a series of Python scripts, which can all be consulted in the project's GitHub repository.[6]

4.1. Fine-tuning Bert models

For the first approach, the 6,014 annotated and curated sentences were used as ground truth to fine-tune multiple Transformer models of the Bert family. We decided to test the following three models:

- *google-bert/bert-base-uncased* (from now on, referred to as *google-bert*), as representative of a large, general-purpose model for English language;
- *LiYuan/amazon-review-sentiment-analysis* (from now on, *LiYuan*), because it was fine-tuned on a similar kind of data (general product reviews) for a similar kind of task (sentiment analysis);
- *JoelVIU/bert-base-uncased-finetuned-amazon_reviews_books* (from now on, *JoelVIU*), because it was finetuned on an even more coherent dataset (Amazon book reviews).

Given the already-discussed issues of unbalance and underrepresentation in the annotation labels, we decided to simplify them by adopting two different strategies:

our tag set and guidelines were addressed during the course of the annotation itself, thanks to the insights we gathered at each new step.

[6] https://github.com/SimoneRebora/CHR2024

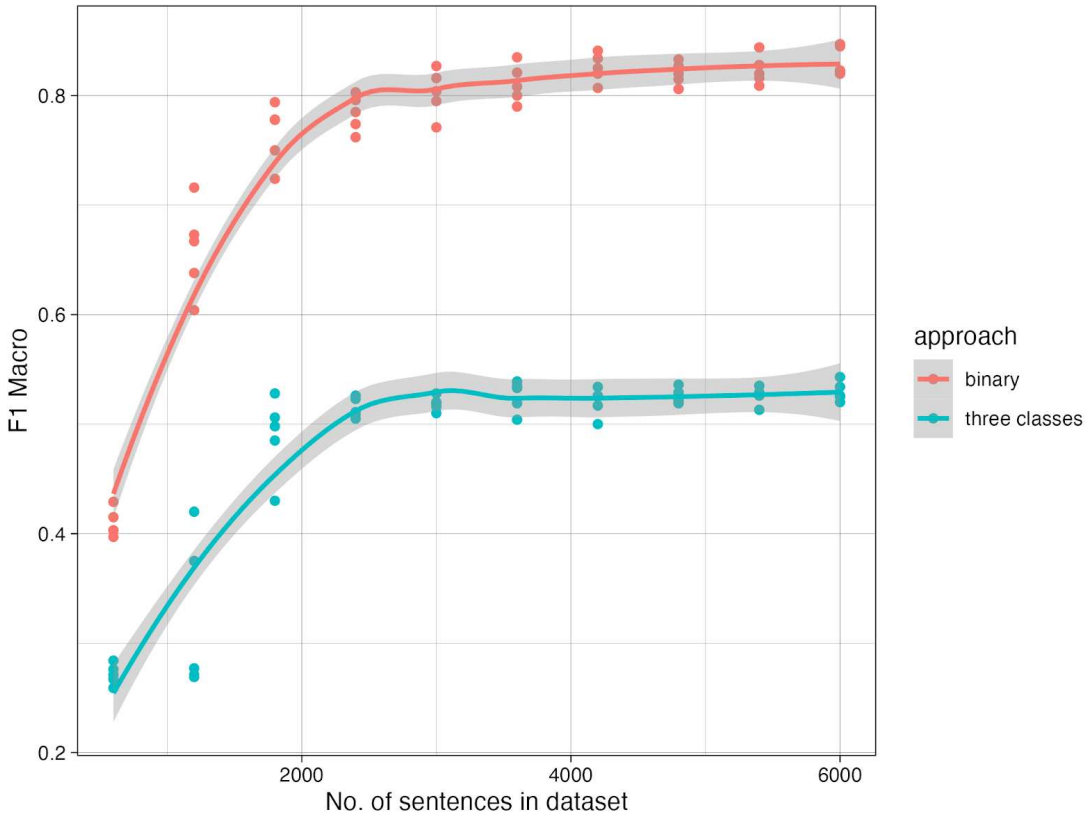

Figure 4: F1-macro scores for the google-bert model, depending on the number of available sentences.

1. we reduced the labels to three classes: 'eval_individual' (under which the three categories related to the impact on the individual reviewer were merged), 'eval_generic' (under which all the evaluation categories were merged), and 'no_val';
2. we further reduced the labels to two classes: 'eval' (under which all the evaluation/impact labels were merged) and 'no_val'.

Testing was first performed via a 5-fold cross validation, to establish which one among the chosen models was the most performant. As shown by Table 2, performance depended on the task, with the *JoelVIU* model slightly outperforming the others for the three-label classification and google-bert producing the best results for the binary classification. However, it should be noted that the three-label setup produced low F1-macro scores because of the substantial failure in classifying the 'eval_individual' sentences (F-1 scores for that label were never higher than 0.144 throughout models/folds), due probably to the very low number of available samples in the dataset (444 sentences, corresponding to 7.4% of the total). In light of these considerations, we decided to use the *google-bert* model as a reference point for our subsequent analyses.

A second level of analysis concerned the fitness of the number of annotated sentences to effectively train the classifier. To address this question, we performed a series of 5-fold cross validations by fine-tuning the model with an increasing number of sentences (from 600 to 6,000,

Table 3
Efficiency report for two classes on the sample dataset

Class	Precision	Recall	F1-score	Support
no_val	0.895	0.933	0.913	854
val	0.817	0.731	0.772	349
Accuracy			0.874	1203
Macro avg	0.856	0.832	0.843	1203
Weighted avg	0.872	0.874	0.872	1203

Table 4
Efficiency report for three classes on the sample dataset

Class	Precision	Recall	F1-score	Support
eval_generic	0.628	0.705	0.664	244
eval_individual	0.000	0.000	0.000	105
no_val	0.871	0.947	0.907	854
Accuracy			0.815	1203
Macro avg	0.500	0.551	0.524	1203
Weighted avg	0.746	0.815	0.779	1203

selected randomly from the dataset). Figure 4 shows how, for both the binary and three-class classification, F1-macro scores reach a plateau at around 3,000 sentences, thus suggesting that the learning threshold for the classifier is substantially lower than the number of annotated sentences.

A third and final level of analysis was chosen in order to get efficiency scores comparable to the ones obtained with Large Language Models (see section 4.2). Here the selection strategy was changed slightly, by randomly choosing entire reviews instead of single sentences, without performing any k-fold cross validation. After testing different configurations, we selected one with 18 reviews in the test set (corresponding precisely to 20% of the sentences in the dataset, with comparable distributions of labels), which produced efficiency scores similar to the ones obtained before (for an overview, see Table 3 and 4).

Given the unsatisfactory performance reached on the three-labels dataset (the model simply ignored the under-represented 'eval_individual' category) we decided to share only the model finetuned for the binary classification of evaluative and non-evaluative sentences, which can now be freely accessed on Hugging Face.[7]

4.2. Instructing GPT 4

For the second approach, the annotation guidelines were used as a starting point to develop system prompts for the GPT-4 Large Language Model. By situating itself in the emergent area of "prompt engineering" [4], the work here became more exploratory, aiming at the identification

[7]https://huggingface.co/GVezzani/literary_evaluation_classifier

of the best technique to instruct the model.

In the first phase of testing, the overall prompting strategy was a simple zero-shot (i.e., instructions plus input), with system prompts (i.e., the instructions) composed by adopting three different approaches:

- The first, defined as *complex*, was a straightforward adaptation of the annotation guidelines in their latest stage. Note that guidelines were loosely structured, with occasional repetitions and multiple addenda;
- The second, defined as *simple*, was a drastic simplification of the annotation guidelines, extracting only the most relevant information and giving it a simpler structure;
- The third, defined as *procedural*, was inspired by the chart in Figure 3 and by the "Tree of Thoughts" (ToT) prompting technique [4], producing a more structured prompt, where the assignment of a label was the result of a set of nested choices.

These three approaches were then combined with three different tagsets:

- The first, defined as *full*, adopted the 7 labels originally used by the annotators (note that this setup was not even tested with Bert models, because of the scarcity of annotations for some labels);
- The second, defined as *3-class*, reduced the labels to the first setup with Bert models: 'eval_individual', 'eval_generic', and 'no_val';
- The third, defined as *binary*, further reduced the labels to the second Bert setup: 'val' and 'no_val'.

For each approach, the system prompt produced with the *full* tagset was then adapted into *3-class* and *binary* by performing the minimum amount possible of modifications, so as to keep the core of the prompt intact. Final result was a set of 9 different system prompts, which can be consulted in Appendix B.

To profit from the ability of GPT-4 to process large amounts of text and to emulate as closely as possible the work of the annotators, we decided to give it as an input entire reviews split into sentences. The user prompt was therefore structured as a .csv file with three columns: 'book_title' (containing the title of the reviewed book), 'sentence_id' (with a numeric identifier of the sentence), and 'sentence' (with the sentence text). At each trial, GPT-4 was prompted 18 times, with the 18 reviews identified with the criteria described in section 4.1.

Requests were processed by using the "gpt-4o-2024-05-13" model (with temperature set to 0 to produce the most deterministic behavior) on the OpenAI API. The whole operation cost a total of 3$. Comparisons of the GPT-4 annotations with the ground truth are shown by Table 5. Note that the adaptation of the *full* tagset to *3-class* and *binary*, and of the *3-class* tagset to *binary* could also be accomplished ex post (i.e., after having performed the annotation with GPT-4).

Overall, GPT-4 efficiency is higher than fine-tuned Bert models for the *3-class* condition (in fact, GPT-4 performs substantially better on the 'eval_individual' label, overcoming the issue of its underrepresentation in the dataset), while it is lower for the *binary* condition. When comparing the different system prompts, the *complex* approach performs slightly better than the *procedural* one, suggesting how a structured prompt may not be necessary to obtain the

Table 5
F1-macro scores for the different system prompts

	full			3-class		binary
adapted ex post to	-	**3-class**	**binary**	-	**binary**	-
complex	0.425	0.634	0.758	0.668	0.803	0.825
simple	0.336	0.531	0.717	0.630	0.801	0.779
procedural	0.386	0.675	0.797	0.641	0.807	0.803

Table 6
Efficiency report for the best-performing prompt (complex_full) on the full tagset

Class	**Precision**	**Recall**	**F1-score**	**Support**
aesthetic	0.312	0.863	0.458	131
generic_val	0.593	0.444	0.508	108
ind_cognitive	0.263	0.333	0.294	15
ind_emotional	0.365	0.613	0.458	75
ind_pragmatic	0.000	0.000	0.000	15
no_val	0.977	0.691	0.809	854
social	0.500	0.400	0.444	5
Accuracy			0.668	1203
Macro avg	0.430	0.478	0.425	1203
Weighted avg	0.809	0.668	0.704	1203

Table 7
Efficiency report for the best-performing prompt (procedural_full, adapted ex-post to 3-class) on the 3-class tagset

Class	**Precision**	**Recall**	**F1-score**	**Support**
eval_generic	0.553	0.807	0.657	244
eval_individual	0.431	0.629	0.512	105
no_val	0.954	0.775	0.855	854
Accuracy			0.769	1203
Macro avg	0.646	0.737	0.675	1203
Weighted avg	0.827	0.769	0.785	1203

highest efficiency. Finally, ex-post tagset adaptation produces mixed results (with the best efficiency for the *3-class* and *binary* tagsets obtained with and without adaptation, respectively), even if in the majority of the cases (6 out of 9) it worsens efficiency. Tables 6, 7, and 8 show more in detail the efficiency of the best-performing setups.

In the second phase of testing, we adopted a "few-shot" strategy by using as a basis the best performing setup (*complex* approach with *binary* tagset). The "few-shot" strategy implies providing not only instructions, but also examples for performing the task. Examples were extracted from the remaining 71 reviews, selecting the ones that showed the proportion of 'val' vs.

Table 8
Efficiency report for the best-performing prompt (complex_binary) on the binary tagset

Class	Precision	Recall	F1-score	Support
no_val	0.962	0.808	0.878	854
val	0.663	0.923	0.771	349
Accuracy			0.841	1203
Macro avg	0.812	0.865	0.825	1203
Weighted avg	0.875	0.841	0.847	1203

'no_val' labels closest to the overall mean in the dataset (i.e. the most "balanced" ones). These reviews were presented to the model together with the curated annotations, before asking it to annotate the new reviews. Three different tests were then performed, with an increasing number of sample reviews:

- two reviews, corresponding to 172 sentences (the whole operation cost 0.66$);
- four reviews, corresponding to 354 sentences (1.21$);
- eight reviews, corresponding to 596 sentences (1.95$).

As prices increased substantially, we decided not to test bigger selections of sample reviews. Also, quite surprisingly, this prompting technique had a detrimental effect on the efficiency of the model, with F1-macro scores decreasing to 0.780 (with two sample reviews), 0.754 (four reviews), and 0.724 (eight reviews).

5. Conclusions

We believe that our work can contribute to the field in several ways. First, the development of a tag set to capture evaluative criteria in unstructured reviews enriches the current search for possible ways to operationalize literary evaluation and study it from an empirical perspective [27, 6].

Furthermore, the application of the aforementioned tag set during the work of annotation revealed some interesting features of the material analyzed. First, let's note that many scholars seem to interpret online book reviewing as animated by a distrust for (if not a full-fledged opposition to) more traditional practices of literary criticism, as embodied by bourdeauian "gatekeepers" [3] such as publishers, professors or jurors of a literary prize. For Franzen, online critics are characterized by "a general distrust of established institutions of aesthetic opinion formation" [9, p. 3]. Even when it is not so clearly thematized, a similar view of today's critical landscape is implicit in many works that aim at unveiling the differences between lay criticism and its traditional counterpart [1, 29]. Last, scholars have interpreted the decline of journalistic criticism as an effect of the rise of internet's rating culture, which, by empowering readers and allowing them to express their own judgements, brings to the slow and inevitable demise of those figures once charged with directing the public's taste [5]. Contrasting such interpretations, the reviewers in our corpus showed no concern for established critical discourse, as demonstrated by the extremely low number of instances of the 'social' label.

The detection of genre-specific evaluative patterns is another interesting finding that deserves to be further investigated in future works. Two interpretations for such result can be hypothesized: first, that the different features of books of different genres [8] somehow 'call for' different evaluative approaches, or, second, that different literary genres have different social constituencies, that is, they are read by different 'kinds of persons' [17], who, in turn, elaborate different evaluative patterns. Unfortunately, our data do not allow us to effectively test either one of these hypotheses.

Last, many researchers are searching for computational ways to analyze the evaluation of books through user-generated online reviews [30, 7]. We contribute to this particular line of research by developing two models for the classification, in online book reviews, of: a) evaluative and non-evaluative sentences and b) generic evaluations, evaluations based on the impact of the book on the reader, and non-evaluative sentences. Furthermore, the comparison between the performance of these models and that of and GPT-4 confirm the findings of recent studies [21, 24], demonstrating the validity of employing LLMs for classification tasks, where they can attain results that are comparable to those of fine-tuned Transformer models. Further research is needed on the reasons for the unsuccess of the "few-shot" prompting technique, which could equally be ascribed to limitations in LLMs or to the complexity of the annotation task. However, this result highlights the importance of applying LLMs to datasets like ours, which could pose new challenges and stimuli for their development.

Acknowledgments

This research was developed in the context of the "Inclusive Humanities" Excellence Project at the Department of Foreign Languages and Literatures of the University of Verona.

References

[1] D. Allington. "'Power to the Reader' or 'Degradation of Literary Taste'? Professional Critics and Amazon Customers as Reviewers of the Inheritance of Loss". In: *Language and Literature: International Journal of Stylistics* 25.3 (2016), pp. 254–278. DOI: 10.1177/0963947016652789.

[2] B. Ashcroft, G. Griffiths, and H. Tiffin. *The Empire Writes Back: Theory and Practice in Post-colonial Literatures*. New accents. London; New York: Routledge, 1989.

[3] P. Bourdieu. *Les Règles de L'Art: Genèse et Structure du Champ Littéraire*. Points Essais. Paris: Éd. du Seuil, 2010.

[4] B. Chen, Z. Zhang, N. Langrené, and S. Zhu. "Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review". In: (2023). DOI: 10.48550/arxiv.2310.14735. URL: https://arxiv.org/abs/2310.14735.

[5] T. Chervel. "18 Die Kritik und ihre Päpste: Rückblick auf ein Genre". In: *Digital Humanities*. Ed. by G. Graf, R. Knackstedt, and K. Petzold. 1st ed. Vol. 2. Bielefeld, Germany: transcript Verlag, 2021, pp. 297–302. DOI: 10.14361/9783839454435-019. URL: https://www.transcript-open.de/doi/10.14361/9783839454435-019.

[6] K. v. Dalen-Oskam. *The Riddle of Literary Quality: A Computational Approach*. Amsterdam: Amsterdam University Press, 2023.

[7] L. De Greve, P. Singh, C. Van Hee, E. Lefever, and G. Martens. "Aspect-Based Sentiment Analysis for German: Analyzing "Talk of Literature" Surrounding Literary Prizes on Social Media". In: *Computational Linguistics in the Netherlands Journal* 11 (2021), pp. 85–104.

[8] A. Fowler. *Kinds of Literature: An Introduction to the Theory of Genres and Modes*. Oxford: Clarendon Press, 1985.

[9] J. Franzen. "Everyone's a Critic: Rezensieren in Zeiten des Ästhetischen Plebiszit". In: *Unterstellte Leseschaften: Tagung, Kulturwissenschaftliches Institut Essen, 29. bis 30. September 2020*. DuEPublico, 2021. DOI: 10.37189/duepublico/74186. URL: https://duepublico2.uni-due.de/receive/duepublico%5C%5Fmods%5C%5F00074186.

[10] E. Gius and J. Jacke. "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis". In: *International Journal of Humanities and Arts Computing* 11.2 (2017), pp. 233–254. DOI: 10.3366/ijhac.2017.0194.

[11] J. Guillory. *Cultural Capital: The Problem of Literary Canon Formation*. First edition, enlarged. Chicago: The University of Chicago Press, 2023. DOI: 10.7208/chicago/9780226830605.001.0001.

[12] R. Heydebrand and S. Winko. "The Qualities of Literatures: A Concept of Literary Evaluation in Pluralistic Societies". In: *The Quality of Literature*. John Benjamins, 2008, pp. 223–239. URL: https://www.jbe-platform.com/content/books/9789027291516-lal.4.16hey.

[13] R. v. Heydebrand and S. Winko. *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. UTB für Wissenschaft Uni-Taschenbücher. Paderborn München: Schöningh, 1996.

[14] R. Kora and A. Mohammed. "A Comprehensive Review on Transformers Models For Text Classification". In: *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. Cairo, Egypt: Ieee, 2023, pp. 1–7. DOI: 10.1109/miucc58832.2023.10278387. URL: https://ieeexplore.ieee.org/document/10278387/.

[15] M. Kuijpers, P. Lendvai, M. Lusetti, S. Rebora, L. Ruh, J. Tadres, T. Ternes, and J. Vogelsanger. "Absorption in Online Reviews of Books: Presenting the English-Language AbsORB Metadata Corpus and Annotation Guidelines". In: *Journal of Open Humanities Data* 9 (2023), p. 13. DOI: 10.5334/johd.116. URL: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.116/.

[16] P. Lendvai, S. Darányi, C. Geng, M. Kuijpers, O. Lopez de Lacalle, J.-C. Mensonides, S. Rebora, and U. Reichel. "Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis. Marseille, France: European Language Resources Association, 2020, pp. 4835–4841. URL: https://aclanthology.org/2020.lrec-1.595.

[17] N. Mark. “Birds of a Feather Sing Together”. In: *Social Forces* 77.2 (1998), p. 453. DOI: 10.2307/3005535.

[18] M. L. McHugh. “Interrater Reliability: The Kappa Statistic”. In: *Biochemia Medica* (2012), pp. 276–282. DOI: 10.11613/bm.2012.031.

[19] F. Pianzola. *Digital Social Reading: Sharing Fiction in the Twenty-First Century*. Cambridge, Massachusetts: The MIT Press, 2025.

[20] W. Piez. “Towards Hermeneutic Markup: An architectural outline”. In: *Digital Humanities Conference*. 2010. URL: https://api.semanticscholar.org/CorpusID:15829935.

[21] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. Robertson, and J. J. V. Bavel. “GPT Is an Effective Tool for Multilingual Psychological Text Analysis”. In: (2023). DOI: 10.31234/osf.io/sekf5. URL: https://osf.io/sekf5.

[22] S. Rebora, P. Boot, F. Pianzola, B. Gasser, J. B. Herrmann, M. Kraxenberger, M. M. Kuijpers, G. Lauer, P. Lendvai, T. C. Messerli, and P. Sorrentino. “Digital Humanities and Digital Social Reading”. In: *Digital Scholarship in the Humanities* 36.Supplement_2 (2021), pp. ii230–ii250. DOI: 10.1093/llc/fqab020.

[23] S. Rebora, M. Kuijpers, and P. Lendvai. “Mining Goodreads. A Digital Humanities Project for the Study of Reading Absorption”. In: (2020). DOI: 10.5281/zenodo.3897251. URL: https://zenodo.org/record/3897251.

[24] S. Rebora, M. L. Lehmann, A. Heumann, W. Ding, and G. Lauer. “Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children’s Literature”. In: *Computational Humanities Research Conference (CHR2023)*. 2023, pp. 333–343.

[25] E. W. Said. *Orientalism*. 1st ed. New York: Pantheon Books, 1978.

[26] M. Salgaro. “Literary Value in the Era of Big Data. Operationalizing Critical Distance in Professional and Non-professional Reviews”. In: *Journal of Cultural Analytics* 7.2 (2022). DOI: 10.22148/001c.36446.

[27] M. Salgaro, P. Sorrentino, L. Gerhard, and A. Jacobs. “How to Measure the Social Prestige of a Nobel Prize in Literature? Development of a Scale Assessing the Literary Value of a Text”. In: *Txt* 1 (2018), pp. 138–48.

[28] C. Van Rees. “How a Literacy Work Becomes a Masterpiece: On the Threefold Selection Practised by Literary Criticism”. In: *Poetics* 12.4–5 (1983), pp. 397–417. DOI: 10.1016/0304-422x(83)90015-3.

[29] M. Verboord. “The Legitimacy of Book Critics in the Age of the Internet and Omnivorousness: Expert Critics, Internet Critics and Peer Critics in Flanders and the Netherlands”. In: *European Sociological Review* 26.6 (2010), pp. 623–637. DOI: 10.1093/esr/jcp039.

[30] K. Wang, X. Liu, and Y. Han. “Exploring Goodreads Reviews for Book Impact Assessment”. In: *Journal of Informetrics* 13.3 (2019), pp. 874–886. DOI: 10.1016/j.joi.2019.07.003.

A. The Tag Set

Here’s a list of the labels in our tag set, each followed by a brief description.

- *no_val*: The sentence does not express an evaluation of the reviewed book
- *generic_val*: Into this category fall all those evaluations related to the work as a whole ("beautiful book," "highly recommended," "an unbearable read," and so on). Also, tag with this category all evaluative sentences that do not fall into any of the following categories.
- *aesthetic*: Any evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...).
- *social*: This value concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that seek to enact (or to reaffirm) a canonization of the judged work, that is, to place it on the roster of 'important' readings.
- *ind_cognitive*: Evaluations regarding the cognitive impact of a book on a reader, the information that the latter extracted from the former or the intellectual stimulation they experienced while reading it.
- *ind_pragmatic*: Evaluations regarding the impact of the book on the reader's life or the existential "lessons" that the latter learned from the former
- *ind_emotional*: Evaluations regarding the emotional impact of the book on the reader, the way the former made the latter feel.

B. The System Prompts

B.1. complex_full

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by "book_title"), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

When the sentence does not express an evaluation of the reviewed book, assign the label "no_val".

When the sentence expresses an evaluation of the reviewed book, you will have to choose between six different labels.

1. "aesthetic": Any evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...).

The following three labels do not refer to features present in the text (such as formal values), but rather to the impact it had on the reader. They are divided into:

2. "ind_pragmatic": Evaluations regarding the impact of the book on the reader's life, the existential "lessons" that the latter learned from the former.

3. "ind_emotional": Evaluation regarding the emotional impact of the book on the reader, the way the former made the latter feel.

4. ”ind_cognitive”: Evaluation regarding the cognitive impact of a book on a reader, the information that the latter extracted from the former or the intellectual stimulation they experienced while reading it.

The last two labels are:

5. ”social”: This value concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that seek to enact (or to reaffirm) a canonization of the judged work, that is, to place it on the roster of ’important’ readings.

6. ”generic_val”: Into this category fall all those evaluations related to the work as a whole (”beautiful book,” ”highly recommended,” ”an unbearable read,” and so on). Also, tag with this category all evaluative sentences that do not fall into any of the above categories.

In assigning one (and only one) of these labels to each sentence, please follow these generic guidelines:

- Work on individual sentences. ONLY in cases where one sentence is incomprehensible without the next, or expresses a concept that necessarily requires continuation in the next, treat the two as a single block and assign them the same label.

- If judgments related to several categories are made in a sentence, tag it with the category that seems the most important to you.

- Judgements regarding other books than the one reviewed, or judgments related to past readings of the same book must be tagged as ”no_val”.

- What does NOT constitute an evaluation: interpretations (”the author wishes to express...” and the like), personal anecdotes (”I first read the book when I was in college”), plot summaries.- ”ind_pragmatic” must contain an explicit reference to the reviewer’s real life or experience.

- ”aesthetic” must have explicit reference to literary art (plot, characters, style).

- An evaluation must specify, explicitly, its object.

- An evaluation may have neutral, or ambiguous (’mixed feelings’) valence.

- Any comparison, resulting in the priority of one over the other, between the book in question and other cultural products is to be considered evaluative.

- All references to story, characters, style, or any other features that relate back to the writing are tagged as aesthetic regardless of the simplicity of the rating. Example: I like the story, I like the characters, etc.

- ”generic_val” consists of all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader.

B.2. complex_3-class

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

When the sentence does not express an evaluation of the reviewed book, assign the label ”no_val”.

When the sentence expresses an evaluation of the reviewed book, you will have to choose between two different labels.

Assign the label ”eval_generic” for any evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...).

The other label, ”eval_individual” does not refer to features present in the text (such as formal values), but rather to the impact it had on the reader. It can be used for:

1. Evaluations regarding the impact of the book on the reader’s life, the existential ”lessons” that the latter learned from the former.

2. Evaluation regarding the emotional impact of the book on the reader, the way the former made the latter feel.

3. Evaluation regarding the cognitive impact of a book on a reader, the information that the latter extracted from the former or the intellectual stimulation they experienced while reading it.

The label ”eval_generic” can also be interpreted in two additional ways:

1. This label concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that seek to enact (or to reaffirm) a canonization of the judged work, that is, to place it on the roster of ’important’ readings.

2. Into this category fall all those evaluations related to the work as a whole (”beautiful book,” ”highly recommended,” ”an unbearable read,” and so on). Also, tag with this category all evaluative sentences that do not fall into any of the above cases.

In assigning one (and only one) of these labels to each sentence, please follow these generic guidelines:

- Work on individual sentences. ONLY in cases where one sentence is incomprehensible without the next, or expresses a concept that necessarily requires continuation in the next, treat the two as a single block and assign them the same label.

- If judgments related to several categories are made in a sentence, tag it with the category that seems the most important to you.

- Judgements regarding other books than the one reviewed, or judgments related to past readings of the same book must be tagged as ”no_val”.

- What does NOT constitute an evaluation: interpretations (”the author wishes to express...” and the like), personal anecdotes (”I first read the book when I was in college”), plot summaries.

- ”eval_individual” must contain an explicit reference to the reviewer’s real life or experience.

- An evaluation must specify, explicitly, its object.

- An evaluation may have neutral, or ambiguous (’mixed feelings’) valence.

- Any comparison, resulting in the priority of one over the other, between the book in question and other cultural products is to be considered evaluative.

- All references to story, characters, style, or any other features that relate back to the writing are tagged as ”eval_generic” regardless of the simplicity of the rating. Example: I like the story, I like the characters, etc.

- ”eval_generic” can consist of all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader.

B.3. complex_binary

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

When the sentence does not express an evaluation of the reviewed book, assign the label ”no_val”.

When the sentence expresses an evaluation of the reviewed book, you will have to use the label ”val”.

Assign this label for any evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...).

Assign the label ”val” also when the evaluation does not refer to features present in the text (such as formal values), but rather to the impact it had on the reader. It can be used for:

1. Evaluations regarding the impact of the book on the reader’s life, the existential ”lessons” that the latter learned from the former.

2. Evaluation regarding the emotional impact of the book on the reader, the way the former made the latter feel.

3. Evaluation regarding the cognitive impact of a book on a reader, the information that the latter extracted from the former or the intellectual stimulation they experienced while reading it.

The label ”val” can also be interpreted in two additional ways:

1. This label concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that seek to enact (or to reaffirm) a canonization of the judged work, that is, to place it on the roster of ’important’ readings.

2. Into this category fall all those evaluations related to the work as a whole (”beautiful book,” ”highly recommended,” ”an unbearable read,” and so on). Also, tag with this category all evaluative sentences that do not fall into any of the above cases.

In assigning one (and only one) of these labels to each sentence, please follow these generic guidelines:

- Work on individual sentences. ONLY in cases where one sentence is incomprehensible without the next, or expresses a concept that necessarily requires continuation in the next, treat the two as a single block and assign them the same label.

- If judgments related to several categories are made in a sentence, tag it with the category that seems the most important to you.

- Judgements regarding other books than the one reviewed, or judgments related to past readings of the same book must be tagged as ”no_val”.

- What does NOT constitute an evaluation: interpretations (”the author wishes to express...” and the like), personal anecdotes (”I first read the book when I was in college”), plot summaries.

- An evaluation must specify, explicitly, its object.

- An evaluation may have neutral, or ambiguous (’mixed feelings’) valence.

- Any comparison, resulting in the priority of one over the other, between the book in question and other cultural products is to be considered evaluative.

- All references to story, characters, style, or any other features that relate back to the writing are tagged as ”val” regardless of the simplicity of the rating. Example: I like the story, I like the characters, etc.

- ”val” can consist of all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader.

B.4. simple_full

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines:

- ”no_val”, when the sentence does not evaluate the reviewed book;

- ”aesthetic”, any evaluation concerning the specifics of literary language, both in its formal and content aspects;

- ”ind_pragmatic”, evaluations regarding the impact of the book on the reader’s life;

- ”ind_emotional”: it is about the value the reader places on the work based on what it made him or her feel. It can range from aspects more related to the book itself, to more intimate and personal issues;

- ”ind_cognitive”: in this category fall all considerations of a book’s ability to teach the reader something or stimulate him intellectually;

- ”social”: this value concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book);

- ”generic_val”: tag with this label all evaluative sentences that do not fall into any of the above categories.

B.5. simple_3-class

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines:

- ”no_val”, when the sentence does not evaluate the reviewed book;

- ”eval_individual”, evaluations regarding the impact of the book on the reader’s life; the value the reader places on the work based on what it made him or her feel (it can range from aspects more related to the book itself, to more intimate and personal issues); all considerations of a book’s ability to teach the reader something or stimulate him intellectually;

- ”eval_generic”, any evaluation concerning the specifics of literary language, both in its formal and content aspects; it also concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book); tag with this label all evaluative sentences that do not fall into any of the above categories.

B.6. simple_binary

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines:

- ”no_val”, when the sentence does not evaluate the reviewed book;

- ”val”, evaluations regarding the impact of the book on the reader’s life; the value the reader places on the work based on what it made him or her feel (it can range from aspects more related to the book itself, to more intimate and personal issues); all considerations of a book’s ability to teach the reader something or stimulate him intellectually; any evaluation concerning the specifics of literary language, both in its formal and content aspects; it also concerns the impact that a book has had not on a single reader, but on a community of readers (references to literary awards, to the popularity of the book); tag with this label all evaluative sentences that do not fall into any of the above categories.

B.7. procedural_full

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

Guideline 1: You will have to treat each sentence as a single unit of meaning. Therefore, you will assign a label based on the sentence alone. Only in cases where one sentence is

incomprehensible without the next, treat the two as a single block and assign them the same label.

Guideline 2: You will have to assign one (and only one) label to each sentence. If several labels can be assigned to one sentence, choose the label that fits best with the sentence.

Guideline 3: Possible labels are: "no_val", "aesthetic", "ind_pragmatic", "ind_emotional", "ind_cognitive", "social", "generic_val". Labels can be grouped into two main categories: non-evaluative sentences and evaluative sentences. When assigning a label to a sentence, you will have to (1) identify the best-fitting category and (2) choose the best-fitting label.

Possible labels are categorized and described here below.

Category 1: Non-evaluative sentences.

A non-evaluative sentence is a sentence that does not express an evaluation of the reviewed book, or that does not explicitly describe the impact it had on the reader.

Category 1, Label 1: "no_val"

Use this label for all non-evaluative sentences. Use it also when the sentence expresses evaluations regarding other books than the one reviewed, or evaluations related to past readings of the reviewed book. Use it also in the case of: interpretations ("the author wishes to express..." and the like), personal anecdotes ("I first read the book when I was in college"), plot summaries.

Category 2: Evaluative sentences.

An evaluative sentence expresses an evaluation of the reviewed book, or it explicitly describes the impact it had on the reader. An evaluative sentence may have positive, negative, or ambiguous ('mixed feelings') valence. An evaluative sentence must specify, explicitly, its object. Any comparison, resulting in the priority of one over the other, between the reviewed book and other cultural products is to be considered as an evaluative sentence.

Category 2, Label 1: "aesthetic"

Use this label when the sentence explicitly expresses an evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...). The evaluation must have explicit reference to literary art (e.g., plot, characters, style).

Category 2, Label 2: "ind_pragmatic"

Use this label when the sentence explicitly describes the impact the reviewed book had on the reader. In particular, use it for sentences regarding the impact of the book on the reader's life, the existential 'lessons' that the reader learned from the book. The sentence must contain an explicit reference to the reviewer's real life or experience.

Category 2, Label 3: "ind_emotional"

Use this label when the sentence explicitly describes the impact the reviewed book had on the reader. In particular, use it for sentences regarding the emotional impact of the book on the reader, the way the former made the latter feel. Still, be careful in not assigning this label to sentences that use an emotional language in a generic and/or metaphorical way (e.g. "I loved this book," "I enjoyed the reading").

Category 2, Label 4: "ind_cognitive"

Use this label when the sentence explicitly describes the impact the reviewed book had on the reader. In particular, use it for sentences regarding the importance of the information that the reader extracted from the book, or the intellectual stimulation the reader experienced while reading it.

Category 2, Label 5: ”social”

Use this label when the sentence evaluates a book not based on the experience of a single reader, but on the experience of a community of readers (e.g., references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that enact (or reaffirm) a canonization of the judged work, placing it on the roster of 'important' readings.

Category 2, Label 6: ”generic_val”

Use this label for all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader (e.g., ”beautiful book,” ”highly recommended,” ”an unbearable read”). Also, tag with this category all evaluative sentences that do not fall into any of the above categories.

B.8. procedural_3-class

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

Guideline 1: You will have to treat each sentence as a single unit of meaning. Therefore, you will assign a label based on the sentence alone. Only in cases where one sentence is incomprehensible without the next, treat the two as a single block and assign them the same label.

Guideline 2: You will have to assign one (and only one) label to each sentence. If several labels can be assigned to one sentence, choose the label that fits best with the sentence.

Guideline 3: Possible labels are: ”no_val”, ”eval_generic”, and ”eval_individual”. They are described here below.

Label 1: ”no_val”

Use this label for all non-evaluative sentences. A non-evaluative sentence is a sentence that does not explicitly express an evaluation of the reviewed book, or that does not explicitly describe the impact it had on the reader. Use it also when the sentence expresses evaluations regarding other books than the one reviewed, or evaluations related to past readings of the reviewed book. Use it also in the case of: interpretations (”the author wishes to express...” and the like), personal anecdotes (”I first read the book when I was in college”), plot summaries.

Label 2: ”eval_individual”

Use this label when the sentence explicitly describes the impact the reviewed book had on the reader.

In particular, use it for sentences regarding the impact of the book on the reader's life, the existential 'lessons' that the reader learned from the book. The sentence must contain an explicit reference to the reviewer's real life or experience.

Use it also for sentences regarding the emotional impact of the book on the reader, the way the former made the latter feel. Still, be careful in not assigning this label to sentences that

use an emotional language in a generic and/or metaphorical way (e.g. ”I loved this book,” ”I enjoyed the reading”).

Finally, use it for sentences regarding the importance of the information that the reader extracted from the book, or the intellectual stimulation the reader experienced while reading it.

Label 3: ”eval_generic”

Use this label for all evaluative sentences. An evaluative sentence is a sentence that explicitly expresses an evaluation of the reviewed book. An evaluative sentence may have positive, negative, or ambiguous (’mixed feelings’) valence. An evaluative sentence must specify, explicitly, its object. Any comparison, resulting in the priority of one over the other, between the reviewed book and other cultural products is to be considered as an evaluative sentence.

Use this label when the sentence explicitly expresses an evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...). The evaluation must have explicit reference to literary art (e.g., plot, characters, style).

In addition, use this label when the sentence evaluates a book not based on the experience of a single reader, but on the experience of a community of readers (e.g., references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that enact (or reaffirm) a canonization of the judged work, placing it on the roster of ’important’ readings.

Finally, use this label for all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader (e.g., ”beautiful book,” ”highly recommended,” ”an unbearable read”). Also, tag with this category all evaluative sentences that do not fall into any of the above cases.

B.9. procedural_binary

You will receive as input a .csv table with the following structure:

book_title,sent_id,sentence

The table includes the review of one book (identified by ”book_title”), split into sentences.

You will have to produce as output another .csv table with the following structure:

sent_id,label

You will assign the label to each sentence (even when the sentence is broken or incomplete) by following these guidelines.

Guideline 1: You will have to treat each sentence as a single unit of meaning. Therefore, you will assign a label based on the sentence alone. Only in cases where one sentence is incomprehensible without the next, treat the two as a single block and assign them the same label.

Guideline 2: You will have to assign one (and only one) label to each sentence. If several labels can be assigned to one sentence, choose the label that fits best with the sentence.

Guideline 3: Possible labels are: ”no_val” and ”val”. They are described here below.

Label 1: ”no_val”

Use this label for all non-evaluative sentences. A non-evaluative sentence is a sentence that does not explicitly express an evaluation of the reviewed book, or that does not explicitly

describe the impact it had on the reader. Use it also when the sentence expresses evaluations regarding other books than the one reviewed, or evaluations related to past readings of the reviewed book. Use it also in the case of: interpretations ("the author wishes to express..." and the like), personal anecdotes ("I first read the book when I was in college"), plot summaries.

Label 2: "val"

Use this label for sentences expressing the impact of the reviewed book on the reader and for all evaluative sentences.

Use this label when the sentence explicitly describes the impact the reviewed book had on the reader. In particular, use it for sentences regarding the impact of the book on the reader's life, the existential 'lessons' that the reader learned from the book. The sentence must contain an explicit reference to the reviewer's real life or experience.

Use it also for sentences regarding the emotional impact of the book on the reader, the way the former made the latter feel. Still, be careful in not assigning this label to sentences that use an emotional language in a generic and/or metaphorical way (e.g. "I loved this book," "I enjoyed the reading").

Finally, use it for sentences regarding the importance of the information that the reader extracted from the book, or the intellectual stimulation the reader experienced while reading it.

Use this label also for evaluative sentences. An evaluative sentence is a sentence that explicitly expresses an evaluation of the reviewed book. An evaluative sentence may have positive, negative, or ambiguous ('mixed feelings') valence. An evaluative sentence must specify, explicitly, its object. Any comparison, resulting in the priority of one over the other, between the reviewed book and other cultural products is to be considered as an evaluative sentence.

Use this label when the sentence explicitly expresses an evaluation concerning the specifics of literary language, both in its formal aspects (use of rhetorical figures, writing style, etc...) and content aspects (character or plot construction, narratological features, etc...). The evaluation must have explicit reference to literary art (e.g., plot, characters, style).

In addition, use this label when the sentence evaluates a book not based on the experience of a single reader, but on the experience of a community of readers (e.g., references to literary awards, to the popularity of the book). Of particular interest here are all those judgments that enact (or reaffirm) a canonization of the judged work, placing it on the roster of 'important' readings.

Finally, use this label for all those statements not accompanied by explanation, i.e., expressions of appreciation not related to specific aspects of the book or to specific effects it had on the reader (e.g., "beautiful book," "highly recommended," "an unbearable read"). Also, tag with this category all evaluative sentences that do not fall into any of the above cases.