Digital Scholarship in the Humanities, 2026, **00**, 1–23
https://doi.org/10.1093/llc/fqag080
Full Paper

OXFORD

What makes a book review compelling? Analyzing informativeness, writing style, and enjoyment

Chiara Alzetta*, , Felice Dell'Orletta, Alessio Miaschi*, and Giulia Venturi*

CNR - Istituto di Linguistica Computazionale "A. Zampolli", ItaliaNLP Lab, via G. Moruzzi 1, Pisa, 56127, Italy
*Corresponding authors. Chiara Alzetta, CNR - Istituto di Linguistica Computazionale "A. Zampolli", ItaliaNLP Lab, via G. Moruzzi 1, Pisa, 56127, Italy. E-mail: chiara.alzetta@cnr.it ; Alessio Miaschi, CNR - Istituto di Linguistica Computazionale "A. Zampolli", ItaliaNLP Lab, via G. Moruzzi 1, Pisa, 56127, Italy. E-mail: alessio.miaschi@cnr.it; Giulia Venturi, CNR - Istituto di Linguistica Computazionale "A. Zampolli", ItaliaNLP Lab, via G. Moruzzi 1, Pisa, 56127, Italy. E-mail: giulia.venturi@cnr.it.

Abstract

Amateur book reviews published on Digital Social Reading (DSR) platforms play a crucial role in sharing reading experiences and capturing reader preferences. However, little attention has been given to the linguistic features that influence the way reading experiences are conveyed and shape readers' perceptions of the aspects discussed in reviews. This study addresses this gap by combining Computational Stylometry and Machine Learning techniques to examine how the linguistic features of book reviews combined with the demographic characteristics of review readers impact the reception of book reviews, focusing in particular on three key aspects that might make a book review compelling, namely informativeness, writing style, and enjoyment. To this aim, we relied on a corpus of Italian Goodreads reviews to investigate how amateur reviewers communicate their reading experience and on a survey to collect human judgments about review reception. Additionally, we investigated the extent to which the review style and the demographic characteristics of review readers can predict the different aspects of review perception. Our findings revealed that linguistic characteristics play a crucial role in shaping reader perceptions and are equally or even more predictive than demographic information in automatic classification models. Nevertheless, while demographic factors such as gender and birth year offer limited utility in forming homogeneous reader groups, reading habits emerged as a relevant factor in identifying shared trends among readers. These insights contribute to a deeper understanding of reader involvement in DSR communities and offer valuable insights for both publishers and review platforms in defining book recommender systems.

Keywords book reviews; human perception; reading experience; stylistic analysis; perception prediction.

1. Introduction

The advent of Digital Social Reading (DSR) has transformed the traditional experience of reading into a dynamic, interactive phenomenon where social interactions play a crucial role in shaping readers' behaviors and preferences (Bronwen 2020). DSR encompasses various practices, among which the act of reviewing books stands out as a particularly insightful activity. Book reviews offer valuable insights into readers' opinions (Kutzner et al. 2021), which publishers commonly analyze to better understand their audience's preferences (Maity et al. 2019; Wang et al. 2019; Dimitrov et al. 2021). Additionally, they provide a deeper understanding of how readers choose their books and articulate their reading experiences (Hajibayova 2019; Rebora et al. 2021). By emphasizing the centrality of readers, studies focused on book reviews enable the identification of collective trends over individual differences, shedding light on broader reading habits and preferences.

The study of reading practices has traditionally been the domain of psychologists, sociologists, linguists, and literary scholars. However, in recent years, computational linguists and computer scientists have also contributed to this research, leveraging computational methodologies to analyze large-scale user-generated data (Kousha et al. 2017; Sabri and Weber 2021; Walsh and Antoniak 2021). This study aligns with these interdisciplinary efforts, focusing on how amateur reviewers communicate multiple aspects of their reading experience and how amateur readers perceive them. Our ultimate goal is to develop a methodology that provides meaningful insights into the writing characteristics that most influence readers' perceptions and make a book review compelling, ultimately guiding readers toward reading the reviewed book. The decision to focus on amateur reviews rather than professional ones stems from the observation that amateur reviews more effectively capture the impact of books on readers. Unlike professional reviews, which are often influenced by academic and editorial conventions, amateur reviews reflect a more spontaneous and personal response to reading (Pianzola 2025).

Previous studies have explored various aspects addressed in reviews. These studies differ in their theoretical premises, the review collections they analyze, and their methodologies, leading to diverse findings (Hegel 2018; Koolen et al. 2020; Rebora et al. 2021; Pianzola 2025). Our research contributes to this debate by introducing a novel methodology that combines Computational Stylometry and Machine Learning techniques to predict how amateur readers perceive multiple aspects of a review's writing style and content. A key methodological feature of our study is the use of a wide range of linguistic features that shape review writing, along with demographic information and readers' reading habits, to examine the impact of these two types of information. To this end, we collected a new dataset of reader ratings through a dedicated survey and analyzed a corpus of Italian book reviews written by amateur reviewers on Goodreads. In line with these premises, we addressed the following research questions:

RQ1: How do amateur book reviewers communicate different aspects of their reading experience in written reviews, and which linguistic features shape their writing style?

RQ2: How do amateur readers perceive various aspects of reviews, and how do demographic characteristics and reading habits influence their perception?

RQ3: To what extent can linguistic writing style features and reader characteristics automatically predict the perception of different aspects of reviews? Which group of features is more predictive, and which aspects are most predictable?

2. Background and motivation

The perception of the reading experience is a multifaceted phenomenon shaped by various linguistic, emotional, and cultural factors. Understanding how readers articulate their experiences in book reviews provides valuable insights into literary reception and consumer behavior. Given the significant influence of product (including book) reviews on purchasing decisions (Chen 2008; Hu et al. 2011; Constantinides and Holleschovsky 2016; Wu et al. 2017; Kutzner et al. 2021), researchers have extensively explored different aspects of review content and writing style to uncover patterns in how readers engage with books and how they express the reading experience.

A crucial area of investigation has been the relationship between review readability and its perceived helpfulness. Studies such as Korfiatis et al. (2008, 2012) have examined how the linguistic complexity and stylistic choices of reviews influence their perceived helpfulness. These studies suggest that book reviews with higher readability tend to be rated as more useful by potential readers. Similarly, Sazzed (2023) focused on readability traits and their correlation with informativeness and review popularity, finding that higher readability often translates into increased engagement and influence. Kutzner et al. (2021), on the other hand, focused on content-related features: the authors manually annotated German book reviews to build a resource useful to investigate the impact of review content on reader perception.

Beyond readability and helpfulness, research has also explored the literary reception of books by analyzing how readers describe their experiences. Boot and Koolen (2020) investigated the lexical choices in book reviews to determine which aspects of a book most significantly impact readers. Their findings highlighted that emotional engagement, narrative structure, aesthetic appreciation, and reflective elements are the most frequently discussed dimensions in book reviews. Their study also proposed a predictive model for identifying sentences in reviews that correspond to these key aspects. Complementary work by Spiteri and Pecoskie (2016) delved into the expression of affect in book reviews, demonstrating that readers convey a rich spectrum of emotions when reflecting on their reading experiences. These studies underscored the deeply personal and subjective nature of literary reception, which is mediated through affective and cognitive responses.

Another critical research direction investigates whether demographic and cultural characteristics of readers influence their literary preferences. Hu et al. (2022) explored how factors such as cultural background and generational differences shape reading choices and review patterns. Among the most widely examined demographic traits is the year of birth, which has been studied in the context of generational reading trends (Chaudhry and Low 2009; Alex 2019; Miranda et al. 2023; Islam and Muna 2024; Prabowo et al. 2025). While some studies suggested generational differences in literary preferences, others, such as Prabowo et al. (2025), indicated that gender and age have minimal impact on reading preferences. These findings challenge assumptions about demographic influences and suggest that individual reader experiences and contextual factors play a more prominent role in shaping literary reception.

Book reviews have also been used as a source of information for predictive modelling. Scofield et al. (2022) employed multiple machine learning algorithms to categorize twenty-four fiction genres based on a dataset of 325 Portuguese book reviews from Goodreads. Similarly, Saraswat (2022) used Amazon book reviews for genre classification, implementing recurrent neural networks to enhance predictive accuracy. Beyond genre classification, Li et al. (2023) investigated the application of ChatGPT in recommendation tasks, including book rating prediction, user rating recommendation, and book summary generation. These studies highlighted the growing role of AI-driven approaches in analyzing user-generated content for personalized recommendations. Another significant line of research focuses on the emotional and cognitive impact of reading experiences as expressed in book reviews. Koolen et al. (2020) developed a model capable of automatically identifying expressions of reading impact in Dutch reviews. Their model categorizes impact into emotional reactions, aesthetic or narrative appreciation, and reflective remarks, demonstrating how computational techniques can capture nuanced reader responses. These studies collectively underscore the potential of computational methods in extracting meaningful insights about reading experience from book reviews, contributing to advancements in both literary analysis and recommendation systems.

To the best of our knowledge, none of these studies have focused on analyzing reader experience using collections of book reviews written by Italian amateur reviewers. One notable exception is Pianzola et al. (2022), who investigated how students and teachers from 211 Italian high schools responded to a highly structured digital social reading activity in terms of engagement intensity and social interaction. This activity involved exchanging tweets to comment on an Italian novel. The authors conducted a social network analysis to examine interactions between participant groups and content analysis to classify the types of content in the tweets, also based on word frequency. Unlike this study, we adopted a user-centered approach, basing our analysis on authentic reviews written by amateur reviewers and exploring how amateur readers perceive various aspects of these reviews in terms of reading impact. Additionally, our study makes an original contribution by combining two main sources of information: linguistic features modelling the writing style of reviews and demographic and reading habit data from amateur readers, investigating their relationship and potential interactions.

3. Book reviews on reading experiences

For this study, we considered two sources of data. The first is a subset of a previously constructed corpus of book reviews written by amateur readers and published on a DSR platform. This corpus, along with its internal composition, is described in Section 3.1. The second source is a newly collected survey composed of five questionnaires, which includes amateur readers' ratings of various aspects of reviews, as detailed in Section 3.2, along with the demographic information gathered from the survey participants (see Section 3.3).

3.1 Book reviews collection

The book reviews rated by the amateur readers were sourced from the GoodReads section of the A Good Review corpus (Alzetta et al. 2024), a collection of around 75,000 book reviews in Italian acquired in June 2022 from Amazon Books and Goodreads for novels spanning various literary fiction genres. To ensure broad genre representation and minimize potential biases arising from participants' personal preferences, each questionnaire of the survey includes reviews from two different books across four distinct genera, resulting in ten books per genre, for a total of forty books.[1]

For each book, we extracted reviews from the A Good Review corpus, limiting the length to 100-120 tokens to minimize the interference of document length with the various characteristics considered in this study. Conversely, to maximize variability across the reviews selected for each book, we considered two surface-level characteristics that serve as proxies for deeper stylistic variations: sentence length and the distribution of emotive lexicon. Accordingly, we selected three reviews for each of the forty books, for a total of 120 reviews, based on the following criteria:

- **Brief sentences review:** The review with the shortest sentences, as determined by the mean number of tokens per sentence.
- **Long sentences review:** The review with the longest sentences, as determined by the mean number of tokens per sentence.
- **Emotive lexicon review:** The review with the highest percentage of words from the A Good Review lexicon (Alzetta et al. 2024), which includes 1,074 manually classified terms (adjectives, nouns, and verbs) extracted from the A Good Review corpus. These terms were manually selected by an expert in literary reception as encompassing emotions, ideas, opinions, and metaphors (e.g., "I have devoured this book," "I was transported by the story"), commonly used to recount the reading experience.[2]

As discussed in the following sections, these three groups serve as a heuristic device to guide further, more refined analyses of variation across reviews. Although based on relatively simple surface-level criteria, they are empirically observable, reproducible, and likely to correspond to deeper differences in stylistic and communicative strategies (see Section 5). For instance, reviews with very short or very long sentences are expected to differ not only in syntactic structure but also, potentially, in perceived engagement. Similarly, reviews rich in emotive and metaphorical expressions may reflect a more personal and emotional tone. This selection strategy also supports our broader aim of investigating whether, and to what extent, these surface-level characteristics of reviews shape reviews' appreciation across multiple dimensions of variations (see Section 6), as well as their role in predicting appreciation levels (see Section 7).

3.2 Human judgments annotation

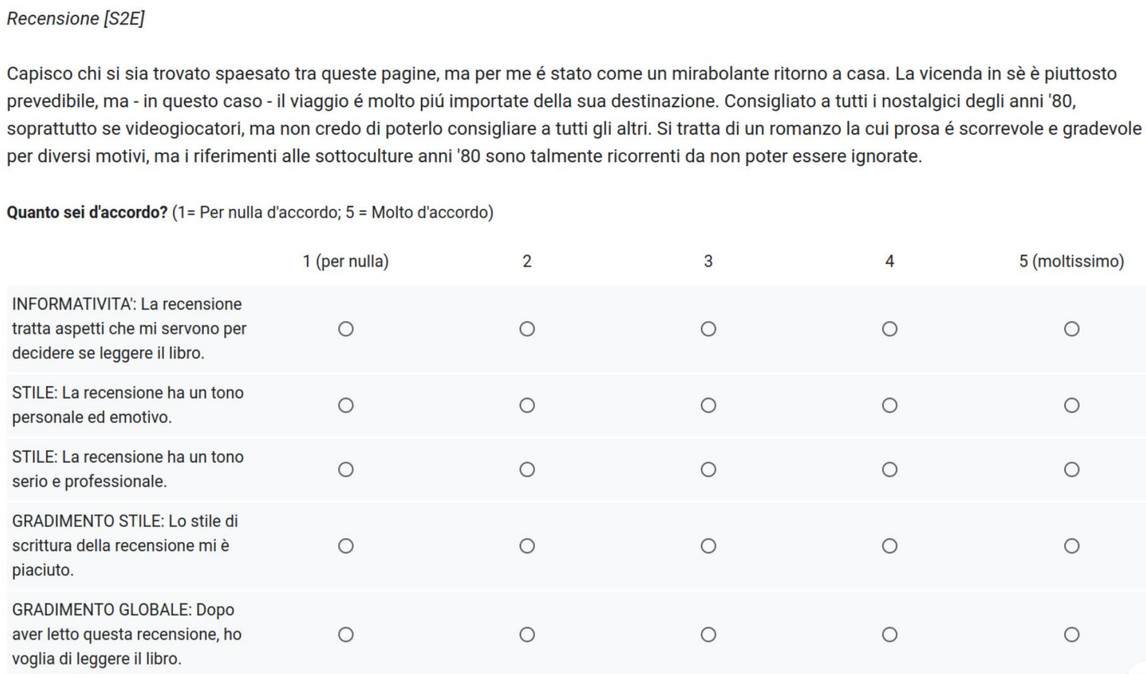
The corpus of 120 reviews was rated by a pool of amateur readers across multiple aspects. Participants in the survey were instructed to imagine themselves as individuals who do not typically enjoy reading for leisure and were visiting a blog to explore reviews of unfamiliar books. This methodological choice was driven by Pianzola (2025, pp. 124–128), who argues that amateur reviews are particularly effective at expressing the emotional and subjective impact of a book, as they are not shaped by academic or editorial constraints. However, the results of previous studies aimed at capturing readers' perceptions of these reviews are often difficult to compare, as they focus on different aspects of review content. Inspired by the methodology adopted by Pianzola (2025), we selected three broad dimensions that capture how amateur reviewers convey their impressions: informativeness, writing style, and enjoyment. We further refined two of these dimensions (writing style and enjoyment) into two sub-aspects each, overall resulting in five evaluative statements. This decision allowed us to explore whether participants perceived dimensions such as personal vs. professional tone or different types of enjoyment as distinct or overlapping. Participants' task thus consisted in rating their agreement with the following five statements, using a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree):

1) **Informativeness**: The review covers aspects that help me decide whether to read the book.
2) **Personal and emotional writing style**: The review has a personal and emotional tone.
3) **Serious and professional writing style**: The review has a serious and professional tone.
4) **Writing style enjoyment**: I enjoyed the writing style of the review.
5) **Global enjoyment**: After reading this review, I am inclined to read the book.

Note that we did not provide formal definitions of concepts such as style, professionalism, or enjoyment, nor were these discussed with participants in advance. This choice was intentional: we aimed to capture spontaneous judgments that reflect individual interpretations, consistent with the informal, blog-like context of the task. While this may introduce variability in how respondents understood specific terms, we view such variation as an inherent part of how amateur readers engage with reviews. This issue is addressed explicitly in Section 6.1.

Figure 1[3] displays the questions presented to participants for one example book review. Since participants were instructed to focus on evaluating the review rather than the book being reviewed, the book title was not disclosed to participants (although they might recognize it from the review content). To ensure participants' workload and time commitment were fair, each questionnaire of the survey comprised twenty-four book reviews. No control questions were included due to the high subjectivity of the task.

The annotation campaign involved 134 subjects who submitted their answers to one of the questionnaires. Participation in the survey was entirely voluntary and uncompensated, with subjects recruited through academic mailing lists and personal acquaintances. The sole participation requirement was being a native Italian speaker, as the survey instructions, book reviews and


Recensione [S2E]

Capisco chi si sia trovato spaesato tra queste pagine, ma per me é stato come un mirabolante ritorno a casa. La vicenda in sè è piuttosto prevedibile, ma - in questo caso - il viaggio é molto piú importate della sua destinazione. Consigliato a tutti i nostalgici degli anni '80, soprattutto se videogiocatori, ma non credo di poterlo consigliare a tutti gli altri. Si tratta di un romanzo la cui prosa é scorrevole e gradevole per diversi motivi, ma i riferimenti alle sottoculture anni '80 sono talmente ricorrenti da non poter essere ignorate.

Quanto sei d'accordo? (1= Per nulla d'accordo; 5 = Molto d'accordo)

	1 (per nulla)	2	3	4	5 (moltissimo)
INFORMATIVITA': La recensione tratta aspetti che mi servono per decidere se leggere il libro.	○	○	○	○	○
STILE: La recensione ha un tono personale ed emotivo.	○	○	○	○	○
STILE: La recensione ha un tono serio e professionale.	○	○	○	○	○
GRADIMENTO STILE: Lo stile di scrittura della recensione mi è piaciuto.	○	○	○	○	○
GRADIMENTO GLOBALE: Dopo aver letto questa recensione, ho voglia di leggere il libro.	○	○	○	○	○

Figure 1. Example of a review and the aspects to be evaluated on the 1 (Completely disagree) to 5 (Totally agree) rating scale.

questions were presented in Italian. To ensure data quality, we excluded responses from subjects who either did not complete the questionnaire or took less than five minutes to finish it, an experimentally determined threshold for adequate completion. After applying these criteria, 100 participants remained, whose answers were considered for the analysis. Specifically, we collected the answers of twenty subjects for each of the five questionnaires, resulting in a final dataset of 12,000 individual responses.

3.3 Demographic profiles of human judges

Demographic information about the participants was collected in the survey before they began the review evaluation. This included strictly demographic details such as birth year, gender, and education level, as well as reading habits, specifically the number of books read per year. In this phase, participants were also informed about the estimated time required to complete the questionnaire, the study's objectives, and terms and conditions.

Demographic data, displayed in Figure 2, allow the profiling of participants, which can be used to investigate the potential relationships between individual characteristics and their responses. The recruited sample is quite balanced in terms of gender, with 56 out of 100 participants identifying as female and only two opting not to disclose their gender. The average age of the full set of participants at the time of the data collection (March 2024) was 38.33 years (standard deviation SD = ±13.60). Both subsets of males and females show a similar average age, with 37.36 years (SD = ±12.89) for males and 38.75 years (SD = ±13.87) for females. To better understand the age distribution, the sample was grouped into social generations, from Baby Boomers to Generation Z.[4] Given the predominance of participants from Generation Y, we further divided this group into two subcategories: Generation Y(80) for those born before 1990 and Generation Y(90) for those born in or after 1990. As shown in the middle graph of Fig. 2, this division results in a more balanced generational distribution. Note also that gender distribution is quite balanced within generations. Regarding educational background, although not illustrated in the figure, all participants reported having at least a high school diploma. Notably, among those with graduate or postgraduate degrees, females (48) outnumbered males (32).

In terms of reading habits, the majority of participants reported reading between four and eleven books per year, especially among subjects belonging to Generation Y, while the number of frequent readers (more than ≥12 books per year) and occasional readers (≤3 books per year) was evenly distributed, also across social generations. Interestingly, these results contrast with national statistics. According to the 2022 annual report on

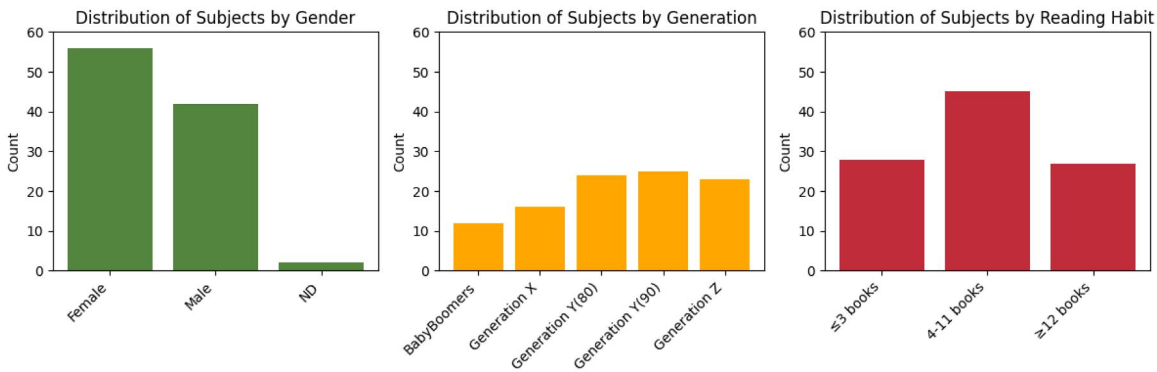


Figure 2. Distribution of subjects by demographic groups based on gender, social generation, and reading habit.

reading habits in Italy, conducted by the National Institute of Statistics (Istat), only 40 per cent of Italians read books regularly. Among them, the largest proportion (44.27 per cent) is classified as Weak Readers (≤3 books per year) (Istat - Istituto Nazionale di Statistica 2023). In contrast, our sample demonstrates a higher engagement with reading: Average Readers (four to eleven books per year) and Strong Readers (≥12 books per year) represent the majority, whereas these groups collectively account for only 21.8 per cent of the Italian population according to the Istat report. This discrepancy was likely due to our recruitment method, which primarily reached individuals with higher levels of education, as well as the nature of our survey, which, being related to reading, was more likely to attract habitual readers.

4. Methodology of analysis

We designed a multilayered analysis of the source data to address the three research questions introduced in Section 1. While the methodological framework is introduced here, the results and discussions are organized as follows: Section 5 addresses RQ1 by analyzing the linguistic characteristics of different types of amateur reviews; Section 6 focuses on RQ2 and examines how readers perceived the collected amateur reviews; and Section 7 answers RQ3 through a series of classification experiments that test the predictive value of different types of features.

More specifically, to answer RQ1, which investigates the writing style of amateur reviews, we examined the linguistic characteristics that define the writing style of amateur book reviews from the A Good Review corpus. Specifically, we automatically extracted a wide range of linguistic features (detailed in Section 4.1), including (morpho-)syntactic and lexical properties, and focused on writing style variations across the three review categories: Brief, Emotive, and Long (see Section 5).

For RQ2, addressing the readers' perception of the reading experience communication, we analyzed participants' responses to the survey questions, examining their distribution in relation to the linguistic characteristics of the reviews and the variability in the ratings provided by the survey participants (see Section 6.1), as well as the demographic traits and reading habits of the survey participants (see Section 6.2).

To address RQ3, focusing on perception prediction, we employed an automatic classification approach based on Machine Learning techniques. As described in Section 4.2, we formulated the task as a binary classification problem and tested various models using different types of linguistic and reader-related information as classification features. Results, reported in Section 7, include a comparative analysis highlighting the most predictive aspects of the reviews and the relative importance of different feature types.

4.1 Linguistic characteristics

The study of the linguistic characteristics modelling the style of the reviews was conducted using the linguistic profiling approach introduced by van Halteren (2004), which uses "large numbers of counts of linguistic features as a text profile." To collect the "large numbers of counts of linguistic features" required for this analysis, we relied on a set of morpho-syntactic properties automatically extracted from the reviews using Profiling-UD (Brunato et al. 2020), a web-based application designed to capture a comprehensive range of linguistic characteristics that characterize language variation within and across texts. The Profiling-UD tool begins by parsing the plain texts to automatically derive their underlying morpho-syntactic structures. This preliminary step results in a linguistically analyzed text following the

Universal Dependencies (UD) annotation schema (de Marneffe et al. 2021), a *de facto* standard for morpho-syntactic corpus annotation based on the dependency syntax paradigm. Subsequently, the tool extracts approximately 150 features from these linguistic representations, as detailed in Table 1.

The selection is inspired by the pioneering work of Douglas Biber and colleagues (see, for example, Biber 1993), who argue that linguistic varieties (or registers, in their terminology) are best characterized by the distributional profiles of recurrent linguistic features, rather than the presence or absence of individual ones. For instance, they noticed that both fiction and textbooks employ nouns and pronouns, but their relative frequencies differ in ways that reflect their distinct communicative goals. Similarly, the stylistic profiles of our review corpus were modeled not only based on sentence length, used as one of the key categorization criteria, but also on the distributional tendencies of a broader set of features. These features, we argue, offer meaningful insights into the communication styles of amateur reviewers, as they are commonly adopted in studies that focus on the *form* of a text (e.g. genre, style, authorship, or readability) rather than on its *content*. This reflects a

Table 1. Linguistic characteristics considered in the study aggregated by group.

Linguistic characteristics	Label
Raw text properties (*RawText*)	
Average document length in sentences	n_sentences
Average sentence length in tokens	Tokens_per_sent
Average word length in characters	Char_per_tok
Vocabulary richness (*LexVariety*)	
New basic italian vocabulary for words and lemmas	NBIV_word, NBIV_lemma
Fundamental/high usage/high availability words of NBIV for words and lemmas	FO_*, AU_*, AD_*
Type/token ratio for the first 100 words and lemmas	TTR_word, TTR_lemma
Morphosyntactic information (*POS*)	
Distribution of part of speech (POS)	[UD pos tag]
Lexical density	Lexical density
Dependency syntactic relations (*SyntDep*)	
Distribution of dependency relations	[UD dependency tag]
Global and local parsed tree structures (*TreeStructure*)	
Average depth of the whole syntactic trees	Tree_depth
Average and maximum dependency link lengths	Link_len-avg, Link_len-max
Average number of prepositional chains per sentence	n. prepositional chains
Total number of prepositional chains and average length	Prep_chain-num, Prep_chain-len
Distribution of prepositional chains by depth	Prep_chain_dist_*
Average clause length	Token_per_clause
Order of elements (*Order*)	
Relative order of subjects and objects with respect to the verb	Subject_pre/post-verbal, Object_pre/post-verbal
Inflectional morphology (*InflectMorpho*)	
Inflectional morphology of lexical verbs and auxiliaries	Verbs/Aux (tense/mood/num, pers/form)
Verbal predicate structure (*VerbPred*)	
Distribution of verbal roots	verbal_root
Distribution of verbal heads per sentence	verbal_head_per_sent
Average verb arity and distribution of verbs by arity	verb_edges-avg, Verb_edges_*
Use of subordination (*Subord*)	
Distribution of principal and subordinate clauses	principal_prop, subordinate_prop
Average length of subordination chains and distribution of chains by depth	subord_chain-len, subord_chains_*
Relative order of subordinate clauses with respect to the principal proposition	subor_pre/post

distinctive aspect of our approach: we relied on general linguistic features that describe the lexical and grammatical structure of a text, rather than on ad hoc features tailored to a specific text type or task.

They can be grouped into nine categories, each corresponding to specific linguistic aspects of a text. These aspects range from simple properties, such as document, sentence, and word length, to characteristics that represent more complex linguistic information of a text, like the distribution of UD Parts-of-Speech (POS), dependency relations, and inflectional properties specific to verbal predicates (e.g., mood, tense, person).[5] More complex features pertain to the global and local syntactic structure of the syntactic tree representing a sentence. These include information about the structure of verbal predicates, the order of nuclear sentence elements (i.e., subject and object) in relation to the verbal predicate, and the use of subordination. Additionally, among the characteristics playing a key role in studying the style of the reviews, we considered the quantitative distribution of certain vocabulary-related features, grouped under the category of Vocabulary Richness. This group encompasses two main types of properties. The first is the Type/Token Ratio (TTR), which measures the ratio between the number of lexical types (lemma) and tokens (word forms) in a text, calculated for both the entire review and its first 100 tokens. The second involves the distribution of word forms and lemmas belonging to the New Basic Italian Vocabulary (NBIV) (De Mauro and Chiari 2016), further classified into three usage categories: fundamental (very frequent words), high usage (frequent words), and high availability (less frequent words referring to everyday objects or actions and thus familiar to most speakers).

4.2 Models and classification setup

The perception prediction task consists of predicting the score assigned by a subject to a survey question. The experiments were performed relying on a Linear Support Vector Machine (LinearSVM) trained and tested by adopting a five-fold cross-validation approach, i.e. train iteratively on 4/5 of the review scores and use the remaining 1/5 as a test set. This ensured the representativeness of the dataset at each iteration. We defined two LinearSVM models, referred to as Profiling and N-grams models, which differ in the text characteristics used to train them. The former takes the set of linguistic characteristics extracted using the Profiling-UD tool as input features. The latter exploits only lexical information since it uses as input feature a contiguous sequence of *n* words acquired from the reviews (i.e. *n*-grams, with *n* equal to 1, 2, and 3). Both models were trained using demographic information of the subject, namely the gender, year of birth, and reading habits. We compared the performance of the above models against a baseline classifier based on the Majority class, which assigns the most frequent class as the output label.

For this experiment, we adopted a binary classification approach. Specifically, we converted the original 1–5 scale judgments into binary (0–1) labels: responses from 1 to 2 were mapped to 0, while responses from 3 to 5 were mapped to 1. Although this approach inevitably collapses the original Likert scale into two poles, which does not fully reflect the continuous judgment scores assigned by participants, it offers a reasonable starting point for addressing a highly subjective task: assessing readers' perception of the appeal of a book review. In this sense, it is intended as a first step toward addressing our third research question: whether demographic information and the linguistic characteristics of book reviews can serve as reliable indicators of the appreciation of amateur reviews. Our intuition is that, if the results of this initial experiment are positive, they may suggest that the considered variables hold reliable predictive value. This would support future extensions of the work involving more fine-grained modeling approaches, such as ordinal classification or regression-based methods, which allow keeping the original continuous judgment scores.

After converting the judgments into binary labels, we observed that the two classes were unbalanced. To address this, we applied a downsampling strategy to the majority class. Specifically, we performed random downsampling 10 times, ran the models on each random sample and computed the average accuracy score of each run. As the variability between accuracy scores was low (SD between ± 0.006 and ± 0.013), we randomly chose one of the samples and used it consistently for all experiments.

5. Stylistic profile of reviews

To address RQ1, we analyzed how amateur reviewers express their reading experience and identified the linguistic features that shape their writing style. We reconstructed the stylistic profile of the reviews, with a focus on identifying the linguistic characteristics that distinguish the Brief, Long, and Emotive review groups. To achieve this, we analyzed the linguistic features acquired using Profiling-UD, applying Kruskal–Wallis, a non-parametric test aimed at assessing whether the median values of a given feature differ significantly across three or more independent groups. The test revealed that approximately 40 per cent (56 out of 143) of the characteristics vary significantly ($P < .05$) between the brief, long, and emotive groups. This finding suggests notable stylistic differences among these groups of reviews. However, while the Kruskal-Wallis test identifies the presence of differences, it does not measure their magnitude. To determine which characteristics exhibit the largest differences, we computed Eta squared (η^2), a measure that quantifies what proportion of variance in a dependent variable (in our case, a linguistic feature) is determined by independent variables (i.e., being part of the brief, long, or emotive review group). η^2 ranges from 0 to 1, with higher values indicating a greater impact of the review groups in driving differences on the values of the linguistic features monitored. Table 2 shows the set of linguistic characteristics with $p \leq 0.001$ and the mean and standard deviation values computed in the full set of reviews and the Brief, Long and Emotive reviews. Features in each group are ordered by decreasing Eta squared (η^2).[6]

Among the characteristics with the highest η^2, two raw-text features stood out: the average length of sentences in terms of tokens (*tokens_per_sent*) and the average length of documents in terms of sentences (*n_sentences*). This result is quite expected and aligns with the design of the three groups of reviews. The two groups containing reviews specifically selected based on sentence length showed the highest and lowest averages: the Brief group included the shortest sentences (average: 13.35 tokens), while the Long group included the longest (average: 37.32 tokens). By contrast, the

Table 2. Stylistic analysis of reviews: for each characteristic with $P \leq .001$, we report the mean and standard deviation values computed in the full set of reviews and the Brief, Long, and Emotive reviews.

Group	Feature	All reviews mean (SD)	Brief mean (SD)	Long mean (SD)	Emotive mean (SD)	η^2
RawText	tokens_per_sent	23.77 (±11.99)	13.35(±2.80)	37.32(±10.02)	20.63(±4.59)	0.77
	n_sentences	5.76 (±2.59)	8.52(±1.85)	3.21(±0.87)	5.55(±1.31)	0.77
	char_per_tok	4.63 (±0.3)	4.52(±0.24)	4.67(±0.36)	4.71(±0.28)	0.07
LexVariety	TTR_lemma	0.25 (±0.32)	0.12(±0.26)	0.40 (±0.32)	0.21(±0.31)	0.12
	TTR_word	0.28 (±0.36)	0.14(±0.30)	0.46(±0.36)	0.25 (±0.35)	0.10
	AD_word	0.1 (±0.03)	0.09(±0.03)	0.11(±0.03)	0.09(±0.03)	0.09
	FO_word	0.72 (±0.07)	0.74(±0.05)	0.69(±0.09)	0.74(±0.07)	0.08
POS	upos_PUNCT	11.62 (±3.34)	13.58(±3.00)	9.72(±2.79)	11.57(±3.09)	0.22
	upos_ADP	11.69 (±3.38)	10.25(±3.11)	13.90(±2.91)	10.92(±3.00)	0.20
	upos_NOUN	15.9 (±3.54)	14.00(±3.39)	17.59(±3.24)	16.11(±3.09)	0.19
	upos_AUX	5.31 (±2.45)	5.98(±2.14)	4.17(±2.32)	5.78(±2.52)	0.11
	upos_PRON	7.15 (±2.65)	8.07(±2.96)	6.02(±1.95)	7.37(±2.57)	0.08
	upos_DET	13.43 (±2.84)	12.51(±2.43)	14.37(±2.93)	13.42(±2.91)	0.06
SyntDep	dep_punct	11.55 (±3.28)	13.45(±2.98)	9.70(±2.71)	11.50(±3.07)	0.21
	dep_case	10.24 (±3.33)	8.84(±2.98)	12.46(±2.93)	9.42(±2.94)	0.21
	dep_nmod	4.72 (±2.24)	4.02(±2.04)	5.99(±2.28)	4.15(±1.86)	0.13
	dep_amod	4.37 (±2.01)	3.54(±1.77)	5.37(±2.31)	4.21(±1.46)	0.12
	dep_aux	2.73 (±1.95)	3.14(±1.77)	1.93(±1.77)	3.11(±2.10)	0.11
	dep_det	12.41 (±2.65)	11.61(±2.35)	13.23(±2.66)	12.40(±2.73)	0.06
TreeStructure	tree-depth	4.76 (±1.89)	3.18(±0.57)	6.77(±1.67)	4.33(±0.95)	0.68
	link_len_avg-max	9.84 (±4.87)	6.07(±1.46)	14.34(±5.16)	9.12(±2.76)	0.59
	link_len-avg	2.6 (±0.37)	2.38(±0.22)	2.79(±0.35)	2.64(±0.39)	0.24
	prep_chain-num	4.15 (±1.9)	3.40(±1.61)	5.36(±1.88)	3.69(±1.62)	0.19
	token_per_clause	8.47 (±2.06)	8.21(±2.00)	9.43(±2.07)	7.76(±1.75)	0.09
InflectMorpho	verbs_Imperfect	4.88 (±10.1)	3.33(±6.74)	2.49(±8.36)	8.83(±13.10)	0.07
	verbs_1st-Singular	25.05 (±30.65)	21.25(±28.65)	17.82(±26.96)	36.08(±33.52)	0.06
VerbPred	verbal_head_per_sent	2.87 (±1.39)	1.69(±0.42)	4.11(±1.30)	2.80(±0.99)	0.63
	verb_edges-avg	2.72 (±0.4)	2.68(±0.42)	2.86(±0.42)	2.62(±0.32)	0.06
	verbal_root	77.4 (±20.19)	70.87(±15.97)	81.55(±22.61)	79.77(±20.23)	0.06
Subord	principal_prop	40.08 (±15.81)	52.18(±13.83)	27.81(±10.54)	40.24(±12.50)	0.44
	subordinate_prop	59.92 (±15.81)	47.82(±13.83)	72.19(±10.54)	59.76(±12.50)	0.44

Features in each group are ordered by decreasing Eta squared (η^2).

Emotive group, selected not for sentence length but for a high concentration of emotive expressions regardless of sentence length, had an intermediate average sentence length (20.63 tokens). This average is in fact very close to that of the entire set of reviews (see the “All reviews” column), suggesting that the Emotive group is not structurally skewed in either direction.

In addition to sentence length, other characteristics help model the writing style of the three considered groups of reviews. These primarily relate to the syntactic structure of the text. Specifically, they model aspects such as global and local syntactic structures (TreeStructure), verbal predicate structures (VerbPred), and the use of subordination (Subord). Consistent with previous analyses, the Long group included sentences with deeper syntactic trees (*tree-depth*), longer syntactic dependency relations (*link_len-max*), and a higher number of verbal heads per sentence (*verbal_head_per_sen*). Additionally, this group also showed a higher percentage of subordinate clauses (*subordinate_prop*).

The three groups of reviews also differ significantly in additional linguistic characteristics with large variance (i.e. with $\eta^2 \geq 0.16$). In particular, reviews in the Long group exhibit a higher percentage of punctuation marks, both in terms of POS distribution (*upos_PUNCT*) and related dependency relations (*dep_punct*). They also included a greater proportion of nouns (*upos_NOUN*) and prepositions, as reflected in their POS tags (*upos_ADP*) and dependency relations typically associated with prepositional complements (*dep_case*). These sentence characteristics may suggest that this group contains more embedded sequences of prepositional complements dependent on a noun (*prep_chain-num*). Notably, all the above reflect aspects of a complex and articulated writing style (Frazier 1985; Gibson 1998; Gildea and Temperley 2010). Interestingly, while characteristics related to lexical variety (*LexVariety*) and verb morphology (*InflectMorpho*) also varied with high statistical significance ($P \leq .001$), they exhibit medium variance (i.e. with $\eta^2 \geq 0.06$). This suggests that while these features differ across review groups, their effect on distinguishing writing styles is less pronounced.

6. Survey answers analysis

This section addresses RQ2, investigating how amateur readers perceive reviews (RQ 2.1), and how demographic characteristics and reading habits influence their perception (RQ 2.2). While RQ 2.2 is addressed in Section 6.2, below we discuss the main results concerned with RQ 2.1.

6.1 Perception of the reading experience communication

6.1.1 Average scores

The detailed analysis of the answers revealed interesting trends across the five aspects evaluated. The analysis focuses on the average scores for each question, calculated by averaging the ratings assigned by all subjects.[7] These scores, presented in Table 3, are grouped by Brief, Long and Emotive reviews. The table also reports the variance and standard deviation of participants’ scores, as well as the inter-rater agreement measured using Krippendorff’s Alpha, which ranges from 1 (full agreement) to 0 (no agreement).

A first result that emerged is the low inter-rater agreement across almost all questions: Krippendorff’s alpha values mostly fell below 0.2, with only Questions 2 and 3 reaching slightly higher levels, indicating weak but not insignificant agreement. We hypothesize that this stems from the distinct characteristics of personal and professional writing styles, which tend to be more consistently recognized by a cohesive group of participants, a point we return to later in the discussion. This outcome aligns with our expectations, given the inherently subjective nature of the task. It also reflects findings in previous work on amateur book reviews, where it has been noted that human judges often struggle to agree on whether a sentence expresses a specific content (Pianzola 2025).

Despite the low agreement, the results of a linear mixed-effects model revealed that the differences in average scores are statistically significant in most cases.[8] Specifically, significant variations were found both across the three groups of reviews (Brief, Long, Emotive) within each question and across the questions themselves. This suggests that, while individual annotators may disagree, there is still a coherent pattern in how the different review types are perceived on average.

The table highlights different scores based on the question. Question 2, which evaluates the perceived personal style of the review, received the highest scores. This suggests that, regardless of the group, the reviews in the survey are generally perceived as having a highly personal style. This interpretation is further corroborated by the results of the linear mixed-effects model: the coefficients associated with Question 2 are all above

Table 3. For each question and review group (RevGroup), we report the average score assigned by the survey participants, the score variance (Var) and standard deviation (SD), and the agreement (Alpha) between participants.

Question	RevGroup	AVG score	Var	SD	Alpha
(1) Informativeness	Brief	2.78	1.36	±1.16	0.121
	Long	3.11	1.48	±1.22	0.191
	Emotive	2.91	1.41	±1.19	0.157
	All reviews	2.93	1.43	±1.20	0.167
(2) Personal style	Brief	3.67	1.31	±1.14	0.270
	Long	3.15	1.49	±1.22	0.325
	Emotive	3.56	1.29	±1.14	0.326
	All reviews	3.46	1.41	±1.19	0.329
(3) Professional style	Brief	2.12	1.19	±1.09	0.284
	Long	2.81	1.62	±1.27	0.365
	Emotive	2.35	1.18	±1.08	0.240
	All reviews	2.43	1.41	±1.19	0.341
(4) Style enjoyment	Brief	2.73	1.50	±1.23	0.140
	Long	2.95	1.51	±1.23	0.177
	Emotive	2.89	1.42	±1.19	0.125
	All reviews	2.86	1.49	±1.22	0.154
(5) Global enjoyment	Brief	2.53	1.48	±1.21	0.125
	Long	2.73	1.58	±1.26	0.155
	Emotive	2.69	1.50	±1.22	0.141
	All reviews	2.65	1.52	±1.23	0.144

0.50 (with confidence intervals, CI, ranging between 0.30 and 0.63), with a coefficient equal to 1 when compared to Question 3, indicating a consistent and substantial difference in how personal style is perceived across review types. This can be attributed to the fact that these are all amateur book reviews, where reviewers likely choose to convey their reading experience through subjective opinions and sensations (Pianzola 2025). In line with these observations, Question 3, capturing the perceived professionalism of the review, records the lowest scores. A more fine-grained analysis shows that personal style is most strongly associated with reviews in the Brief group, while reviews in the Long group tend to be perceived as more professional. These tendencies are statistically supported: the linear mixed-effects model yields a coefficient of 0.52 between Brief and Long reviews for Question 2 (with CI ranging between 0.40 and 0.63), and 0.69 for the same comparison in Question 3 (with CI ranging between 0.58 and 0.80).

The second highest average score was recorded for Question 1 (2.93), indicating that the reviews were perceived as moderately informative. This question also showed the most pronounced statistical difference when compared to Questions 2 and 3, with a linear mixed-effects model coefficient of approximately 0.50 (CIs between 0.43 and 0.59) in both cases. These results suggest that ratings related to informativeness differ markedly from those concerning the writing style of the review. Within this question, differences among the review groups also emerge: the Long group received the highest average score (3.11), with a statistically significant coefficient of 0.33 (CIs between 0.20 and 0.44) compared to the Brief group.

Interestingly, the two questions related to the enjoyment of the review (Questions 4 and 5) yielded slightly different average scores (2.86 and 2.65, respectively), and both showed their highest divergence with respect to Question 2, with coefficients of 0.60 (CI between 0.53 and 0.66) for Question 4 and 0.81 (CI between 0.74 and 0.87) for Question 5. This indicates that, while participants tended to appreciate the style of the reviews more than the reviews as a whole, their perception of enjoyment diverged most strongly from their assessments of personal tone. In both cases, differences across review groups were less pronounced than for other questions, yet still remained statistically significant.

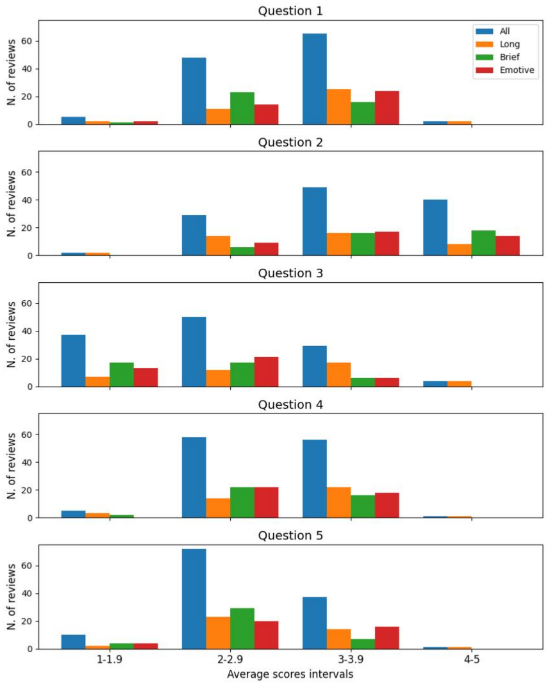


Figure 3. Distribution of all reviews (All) across average scores assigned by the survey participants for each question, also considering the distribution by review types (Long, Brief, and Emotive).

The results reported in Table 3 are complemented by Fig. 3, which illustrates the distribution of reviews across the average scores assigned by survey participants on a 1–5 scale. For comparison, we report the number of reviews for each average score interval, considering the full set of 120 judged reviews (All) as well as the three subgroups Brief, Long, and Emotive (each containing forty reviews). In particular, this visualization helps illustrate to what extent the surface-level characteristics used to define the three groups may correspond to different distributions of reader judgments. Several patterns emerge. Reviews with short sentences were often perceived as more personal, though this writing style was also associated with a lack of professionalism and seriousness. In fact, for Question 2, eighty-nine reviews received high scores (i.e., 3 or above), with a slightly higher presence of Brief reviews (thirty-four), particularly in the highest score interval (eighteen). Conversely, for Question 3, most reviews (eighty-seven) received low scores (i.e., below 2.9), especially those in the Brief group (thirty-four). Additionally, reviews containing long sentences were perceived as more informative than those with short sentences. For Question 1, sixty-seven reviews received scores between 3 and 5, indicating that many

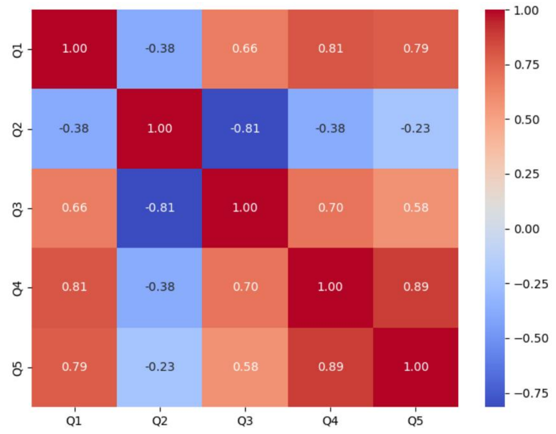


Figure 4. Pearson correlation among the average scores assigned to each of the five questions, with $P \leq .05$. Only the Q2–Q5 pair falls below significance under the Bonferroni correction.

participants found them informative, with twenty-seven of these belonging to the Long group. Reviews with long sentences were also associated with a professional and serious tone. In particular, twenty-one of the reviews rated as highly professional (i.e., with scores between 3 and 5 for Question 3) were in the Long group. For Questions 4 and 5, we observe similar score distributions, with a slight preference for reviews with long sentences, whose overall writing style (Question 4) was more appreciated than that of reviews with short sentences. Finally, the use of emotive lexicon mostly influenced the perception of reviews with a personal style, as reflected in the number of reviews receiving high scores for Question 2.

6.1.2 Correlations among the answers to the five questions

These observations are supported by the correlation scores reported in Fig. 4. The heatmap displays the Pearson correlations among the average scores assigned to the questions. All correlations are statistically significant (with $P \leq .05$), indicating that the survey questions are interrelated. To account for the multiple comparisons problem, we applied both Bonferroni and False Discovery Rate (FDR) corrections to the associated P-values. The results confirm the robustness of the correlations: all pairs remained significant under FDR, and only one pair fell below the significance threshold under the Bonferroni correction. This is the case for the Q2–Q5 pair, which also had the lowest original Pearson coefficient (−0.23).

Consistent with the previous results, we observe a strong negative correlation between Questions 2 and 3. Together with the large mixed-effects coefficients for this pair, this finding indicates that survey participants treat personal and professional tone as opposite ends of a single stylistic dimension rather than as independent traits. Thus, when subjects identify a review as having a personal style, they also coherently assign a low score to the question on the professional style. Notably, this trend supports our initial assumption when designing the survey: namely, that the broad aspect of *writing style*, commonly referenced in prior studies on amateur review content (Pianzola 2025), could be meaningfully decomposed into two opposing poles of the same stylistic dimension.

A second notable pattern emerges for Question 1. Its correlation is positive (0.66) with Question 3 and weakly negative with Question 2 (−0.38), suggesting that participants tend to associate informativeness with a more professional, and less overtly personal, style. This is consistent with the intuition that informativeness is more often associated with reviews perceived as professional rather than emotionally expressive. At the same time, we find a strong positive correlation between informativeness and both dimensions of enjoyment (Questions 4 and 5). This suggests that when participants agree with the statement in Question 1 ("The review covers aspects that help me decide whether to read the book"), they tend to assign similarly high scores to Question 4 ("I enjoyed the writing style of the review") and Question 5 ("After reading this review, I feel like reading the book"). Finally, the very strong correlation between Questions 4 and 5 (0.89) suggests that, as expected, writing style and overall enjoyment are perceived as closely related expressions of the same broader dimension of "enjoyment."

As expected, the strongest positive correlation is observed between responses to Questions 4 and 5, suggesting that enjoyment of a review's writing style is closely linked to overall enjoyment of the review. Consistent with previous observations, the second strongest correlation is between Questions 2 and 3. In this case, a strong negative correlation confirms that, when subjects identify a review as having a personal style, they also coherently assign a low score to the question on the professional style. The correlation results also complement the findings on the distribution of reviews based on scores assigned to Question 1. Despite differences in distribution, we observe a strong positive correlation between the scores assigned to this question and those assigned to Questions 4 and 5. This suggests that informativeness influences the perception of both stylistic and overall enjoyment of the review. Interestingly, the correlations between Question 1 and the perceptions of a professional writing style (Question 3) are lower and even negative when compared with perceptions of personal style (Question 2). This is expected and confirms that aspects revealing the informativeness of the reviews tend to be related to the professional tone of the reviews, while scarcely related to the personal style of the review. Taken together, these results suggest that participants, even with different personal interpretations of the questions eliciting review perception, rely on shared underlying intuitions that produce coherent patterns at the aggregated level.

6.1.3 Stylistic profile and human evaluation scores

The analysis of participants' responses is further enriched by an investigation into the relationship between the assigned scores and the linguistic characteristics used to profile review styles (see Section 5). Specifically, the analysis examines the linguistic features that show a significant positive or negative Pearson correlation ($P \leq .05$) with the assigned scores, aiming to determine whether the writing style of reviews implicitly influences human judgments. To account for the multiple comparisons involved in the correlation process, we applied a False Discovery Rate (FDR) correction.[9] After the correction, the number of significant correlations naturally decreases, but the overall pattern remains consistent. Specifically, the percentage of linguistic features, including the use of emotive lexicon identified by Alzetta et al. (2024), significantly correlated with participants' scores is 28.87 per cent for Question 1, 32.39 per cent for Question 2, 37.32 per cent for Question 3, 17.61 per cent for Question 4, and 4.93 per cent for Question 5. While these percentages are not particularly high, they remain noteworthy also after the FDR correction, especially given that participants were not explicitly asked to consider linguistic characteristics when responding to the questions. Interestingly, the highest number of correlated linguistic characteristics is observed for Questions 2 and 3, which directly relate to the perception of writing style, while the lowest percentages are found for Questions 4 and 5, which focus on the enjoyment of reviews, an aspect of perception that is only marginally related to specific features of writing style.

Despite the amount of FDR-corrected correlated characteristics, the ranking of the most correlated ones varies depending on the focus of the question, as it can be seen in SM3. Each question's ranking includes linguistic characteristics that capture different and sometimes heterogeneous aspects of writing style, without converging on a single communicative function that the question might be expected to elicit. In other words, while individual features correlate with participants' judgments, no consistent pattern emerges across questions.

6.2 Demographic-based analysis on review perception

Based on previous analyses, we observe a substantial variation among survey participants, as indicated by the low levels of agreement in their responses. This led to the hypothesis that there might be a higher level of homogeneity among participants who share certain demographic traits. To investigate whether demographic characteristics influence reading experiences and preferences (RQ 2.2), we analyzed the distribution of survey scores within homogeneous groups of participants. Specifically, we categorized participants based on gender, social generation, and reading habits, following the categorization proposed in Fig. 2.

Table 4 presents the demographic groups' average scores assigned to each question, along with their inter-annotator agreement.[10] Overall, the average scores for these groups closely align with those of the full participant set, including internal variance and standard deviation (see Table 3). However, some differences emerge, since when applying a linear mixed-effect model, we noticed that these differences are statistically significant only in a few cases, and no consistent pattern is observed across all five questions.[11]

Notably, for Question 3 (professional style) and Question 4 (global style), strong readers (≥ 12 books/year) gave significantly lower ratings, with model coefficients of -0.25 (CI between -0.37 and -0.11) and -0.23 (CI between -0.36 and -0.09), respectively. This may suggest a more critical attitude among frequent readers. Interestingly, the highest agreement overall is achieved by those who read the least (≤ 3 books per year group), possibly indicating a more uniform perspective among occasional readers, who may rely on broader or more general criteria when assessing reviews.

Regarding the gender-based groups, male and female scores appear closely aligned, suggesting minimal differences in reading experiences and preferences. However, a linear mixed-effect model reveals that in most cases (excluding Question 2) there are statistically significant differences between genders: female participants tend to give slightly higher scores overall. This effect is particularly evident for Question 1 (Informativeness), with a coefficient of 0.23 (CI between 0.12 and 0.32).

The analysis of social generations reveals that younger participants (Generations Y(90) and Z) tend to assign higher scores when evaluating reviews, whereas older generations (Baby Boomers and X) exhibit the opposite trend. This pattern is further supported by agreement levels, as the highest alpha scores are observed among either the youngest or the oldest participants. Moreover, a linear mixed-effect model analysis shows statistically significant differences in several cases. Notably, for Question 1, Baby Boomers gave markedly lower scores compared to Generation Y_90 participants (2.83 vs. 3.11 on average), with a model coefficient of -0.35 (CI between -0.53 and -0.17). One possible interpretation is that younger readers, who are more accustomed to digital content and online reviews, are more enthusiastic in their evaluations. In contrast, older generations, potentially having more experience with traditional literary criticism or professional reviews, are more reserved in assigning high scores.

These findings highlight that demographic factors only subtly influence reading preferences and review evaluation tendencies. Interestingly, our results align with the existing research on Digital Social Reading, which acknowledges that relationships among "members of online reading communities are based on a principle of affinity, rather than geographical, educational, or demographic traits" (Pianzola 2025: 47). While broad trends emerge from our analysis, the overall similarities across groups seem to suggest that other characteristics of readers, such as spontaneous common interests or similar life experiences across groups of readers (Pianzola 2025: 48–49), may play a more significant role in shaping shared evaluative judgments than demographic characteristics alone. We leave further investigation of this aspect for future research.

6.2.1 Global enjoyment across demographic groups

Despite our earlier observation about the weak impact of demographic characteristics in identifying collective trends in review perception, we noted that the statistical tests used do not reveal whether a subset of reviews is consistently liked (or disliked) by all members of a demographic group, an aspect potentially relevant for predictive modeling scenarios, such as book recommendation tasks. To explore this hypothesis, we focused on the responses to Question 5, which evaluates readers' Global Enjoyment of amateur reviews. We adopted an overlap-based analysis inspired by evaluation metrics used in recommendation systems and information retrieval. These metrics, such as those based on the "Find Good Items" principle, assess how often a relevant item appears among the top-*k* ranked results [10.1145/963770.963772]. Our methodology was structured as follows. First, we ranked the whole set of 120 reviews in descending order based on their average score for Question 5. Second, we selected the top quartile (thirty most-liked) and bottom quartile (thirty least-liked). Then, for each demographic group, we counted how many reviews appeared in the same quartile for every category (e.g. for the Gender group, reviews that

Table 4. For each question and demographic group, we report the average score assigned by the survey participants to the entire set of reviews, the score variance (Var) and standard deviation (SD), and the agreement (Alpha) between participants.

Question	DemGroup	AVG score	Var	SD	Alpha
Q1	Female	3.05	1.49	±1.22	0.157
	Male	2.83	1.32	±1.15	0.158
	BabyBoomers	2.83	1.63	±1.28	0.017
	Generation X	2.99	1.22	±1.10	0.241
	Generation Y(80)	2.99	1.35	±1.16	0.157
	Generation Y(90)	3.11	1.24	±1.11	0.208
	Generation Z	2.88	1.34	±1.16	0.105
	≤3 books	2.98	1.51	±1.23	0.213
	4–11 books	2.95	1.49	±1.22	0.138
	≥12 books	2.91	1.29	±1.14	0.156
Q2	Female	3.41	1.57	±1.25	0.348
	Male	3.54	1.18	±1.08	0.279
	BabyBoomers	3.27	1.08	±1.04	0.273
	Generation X	3.28	1.31	±1.14	0.212
	Generation Y(80)	3.45	1.54	±1.24	0.421
	Generation Y(90)	3.56	1.45	±1.20	0.323
	Generation Z	3.45	1.41	±1.19	0.243
	≤3 books	3.39	1.53	±1.24	0.365
	4–11 books	3.57	1.52	±1.23	0.344
	≥12 books	3.39	1.22	±1.10	0.294
Q3	Female	2.51	1.40	±1.18	0.351
	Male	2.38	1.39	±1.18	0.295
	BabyBoomers	2.36	1.31	±1.15	0.246
	Generation X	2.43	1.30	±1.14	0.340
	Generation Y(80)	2.40	1.38	±1.18	0.333
	Generation Y(90)	2.48	1.32	±1.15	0.343
	Generation Z	2.54	1.49	±1.22	0.417
	≤3 books	2.59	1.58	±1.26	0.355
	4-11 books	2.37	1.4	±1.18	0.309
	≥12 books	2.34	1.24	±1.11	0.362
Q4	Female	2.96	1.54	±1.24	0.138
	Male	2.79	1.37	±1.17	0.164
	BabyBoomers	2.95	1.68	±1.30	0.065
	Generation X	2.76	1.29	±1.14	0.205
	Generation Y(80)	2.93	1.36	±1.17	0.135
	Generation Y(90)	2.95	1.36	±1.17	0.070
	Generation Z	2.87	1.52	±1.23	0.213
	≤3 books	2.96	1.50	±1.23	0.140
	4–11 books	2.88	1.51	±1.23	0.125
	≥12 books	2.76	1.41	±1.19	0.136
Q5	Female	2.73	1.56	±1.25	0.135
	Male	2.58	1.49	±1.22	0.133
	BabyBoomers	2.56	1.75	±1.32	0.004
	Generation X	2.66	1.56	±1.25	0.246
	Generation Y(80)	2.73	1.37	±1.17	0.182
	Generation Y(90)	2.72	1.41	±1.19	0.123
	Generation Z	2.65	1.52	±1.23	0.094

(continued)

Table 4. (continued)

Question	DemGroup	AVG score	Var	SD	Alpha
	≤3 books	2.60	1.55	±1.25	0.146
	4–11 books	2.69	1.51	±1.23	0.125
	≥12 books	2.64	1.49	±1.22	0.098

Table 5. Number of reviews (in per cent between parenthesis) in the quartiles of the most and least appreciated review shared by all categories within each demographic group.

DemGroup	Appreciated	Not appreciated
Gender	16 (53.33%)	17 (56.67%)
Generation	0 (0%)	2 (6.67%)
Reading habit	9 (30%)	8 (26.67%)

appeared in the same quartile for both females and males).

The results reported in Table 5 show that very few reviews were consistently appreciated or disliked by all members of any given group. As expected, females and males exhibit a higher level of consensus, with a greater number of reviews commonly appreciated and not appreciated, accounting for approximately 55 per cent of the reviews in both quartiles. However, this is possibly due to the fact that the gender-based grouping consists of only two categories. Interestingly, no review was consistently appreciated by all groups of survey participants when aggregated for age. This seems to suggest that age-based categorization groups individuals with the least affinity in terms of review enjoyment. In contrast, reading habits appear to cluster individuals with a higher level of shared preferences.

Building on these findings, a complementary research question arises: how do demographic traits influence the distribution of appreciated and not appreciated reviews across the three review groups, i.e. Brief, Long, and Emotive, regardless of the overlap among group members? For example, do both females and males appreciate reviews in the Long group or do females tend to dislike reviews in this group? To address this, we analyzed all reviews in the first and fourth quartiles, as above, without restricting our focus to only those reviews that overlapped between categories within each demographic group. Results are reported in Fig. 5, which illustrates, for each demographic group, the distribution of reviews in the Brief, Long, and Emotive groups within the two quartiles.

Quite interestingly, this complementary analysis suggests that demographic factors may shape the types of reviews readers tend to appreciate or dislike. Specifically, it reveals that, across all three groups (gender, generation, and reading habits), Emotive reviews appear far more frequently among the least appreciated review types. This pattern is particularly pronounced among older participants (Generation X), female respondents, and less frequent readers (≤3 books per year). Conversely, Brief and Long reviews are consistently more appreciated across categories, with only minor variations. A more notable difference emerges among moderately frequent readers (those who read between four and eleven books per year), who show a marked preference for Brief reviews.

7. Predicting aspects and enjoyment of reviews

This section provides an in-depth analysis of the experiments conducted to address RQ3, focusing on predicting the perception of five aspects that make book reviews compelling. Following our proposed methodology, we frame this task as a binary classification problem. The classification results are reported in Table 6. As outlined in Section 4.2, we experimented with two classification models, Profiling and N-grams, differing in the text characteristics used as input features. All results are computed in terms of Accuracy, defined as the ratio of review scores correctly classified with a 0 or 1 label out of the total number of classified scores, for each question.

Our first observation is that the two models incorporating linguistic information, either stylistic, as in the Profiling model, or solely lexical, as in the N-grams model, outperform the majority class baseline, which achieves an accuracy of 0.5 on our balanced set. This suggests that, despite the low inter-annotator agreement in the questionnaire responses, there are consistent patterns across the ratings that can be successfully exploited for prediction.

Furthermore, we notice that all models, including the baseline, face the highest challenges when predicting the scores of Questions 4 and 5, while Question 3 is the best-predicted one. The lower performance on Questions 4 and 5 is quite expected, as these relate to the enjoyment of the review, a highly subjective aspect of evaluation that

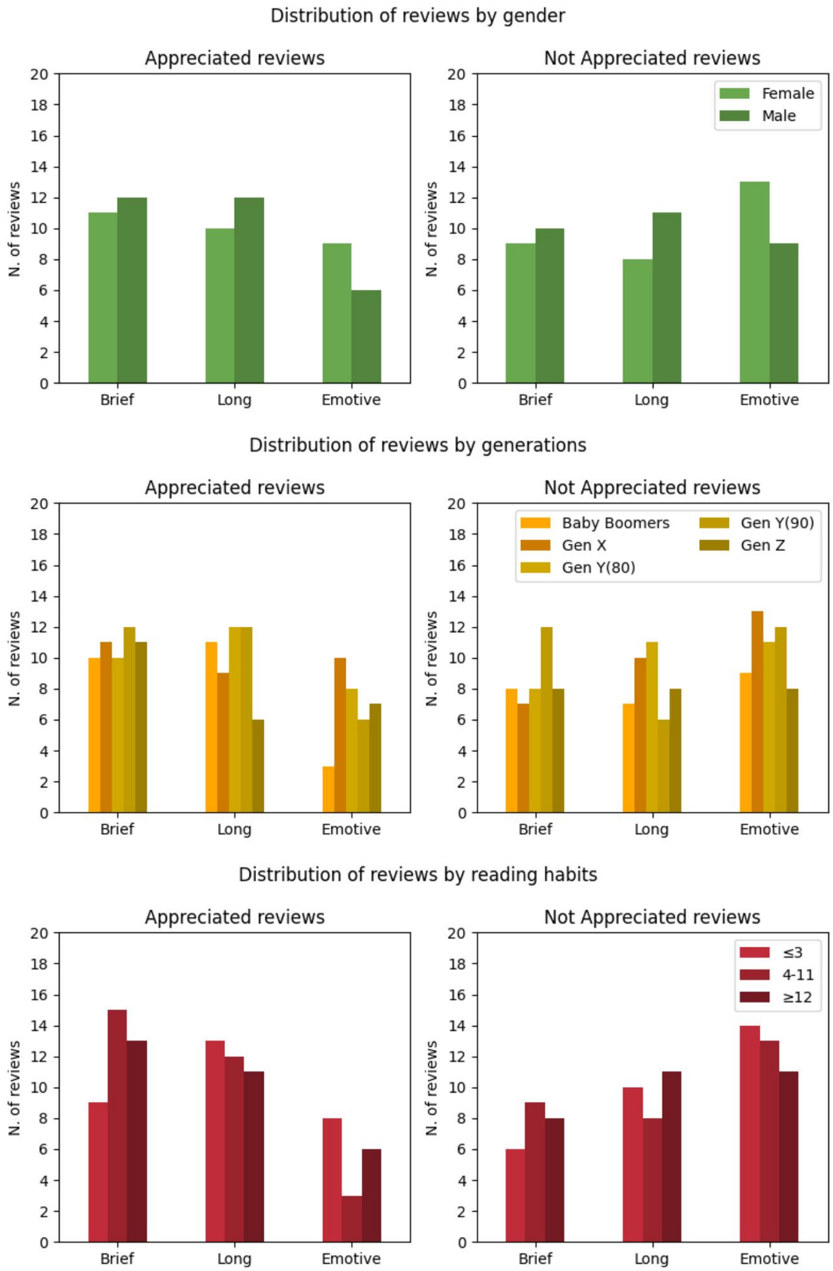


Figure 5. Distribution of appreciated and not appreciated reviews across the Long, Brief, and Emotive review groups, analyzed by demographic categories: gender, social generation, and reading habits.

potentially encompasses multiple characteristics of the review, thus making the automatic prediction of scores more challenging. In contrast, the highest accuracy observed for Questions 1 to 3 indicates that linguistic and lexical characteristics are essential for predicting the readers' rating of informativeness and style.

7.1 Prediction analysis

The overall classification results, while not particularly high, are not surprising given, as previously observed, the low inter-annotator agreement in the questionnaire responses, a factor that further reflects the inherent

Table 6. Accuracy of the binary classification models using linguistic characteristics (Profiling) and *n*-grams (N-grams) as features.

Question	Profiling	N-grams
Q1—Informativeness	0.641	0.641
Q2—Personal style	0.669	0.668
Q3—Professional style	0.721	0.720
Q4—Style enjoyment	0.618	0.626
Q5—Global enjoyment	0.616	0.620

subjectivity of the task. However, our primary goal is not to maximize classification accuracy, but rather to investigate the role played by different sets of features in shaping model predictions. To this end, we analyze the predictions produced by the Profiling and N-grams classification models not only in terms of overall accuracy but also by comparing their error patterns. The results indicate that the predictions overlap quite consistently, with only about 2 per cent of the scores misclassified by either the Profiling or N-grams model in all questions. Although both models exhibit similar predictions, further investigation is needed to identify the linguistic characteristics of the reviews and which demographic information about the human raters is primarily used by each model to achieve the observed accuracies.

7.1.1 Feature importance

To investigate the relevance of features for the prediction models, we ranked all features used by the two systems based on their importance. Our aim is to investigate whether there is a relationship between the importance of the features used by the two classifiers and whether this relationship changes across the prediction of different review aspects. The heatmaps in Fig. 6 report the results of this analysis. They illustrate the Spearman correlation scores between feature rankings for each question in the Profiling and N-grams models. While the two heatmaps display different correlation values, they reveal the same underlying patterns: the strongest correlations consistently emerge between the same pairs of questions, specifically Q4 and Q5, both assessing enjoyment, and Q1 and Q5, capturing informativeness and global enjoyment. These findings align with the trends observed in Fig. 4 for readers' scores, reinforcing two key insights: first, that appreciating a review's style increases the likelihood of reading the book it discusses; and second, that a reader's willingness to read a book is closely linked to the level of informativeness conveyed in the review.

The second goal concerns the analysis of the potentially different roles played by the two main groups of

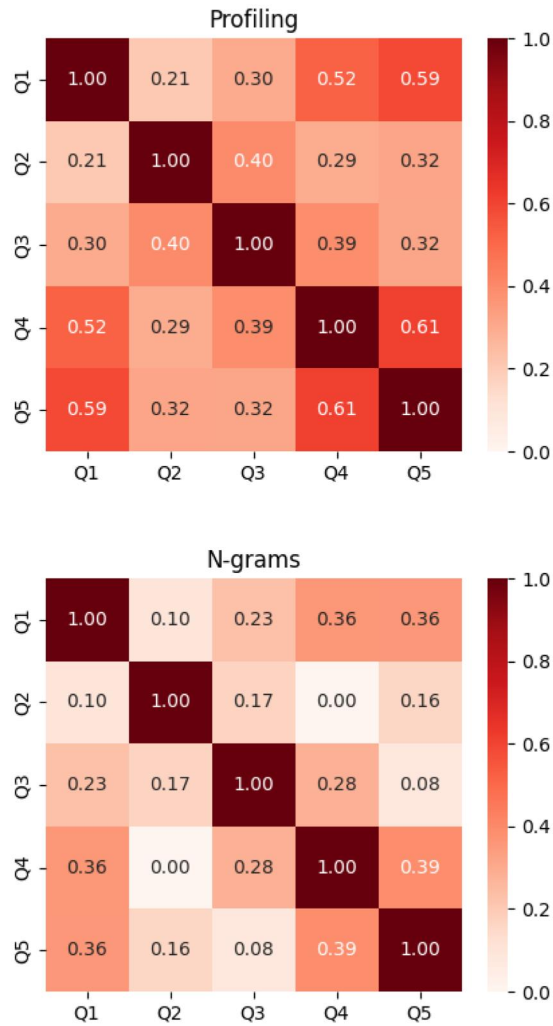


Figure 6. Correlation of feature rankings for the Profiling and N-grams classification models, with $P \leq .05$ except for Q2/Q4.

classification features: one comprising either the Profiling-UD characteristics or the *n*-grams, and the other including demographic information about the survey participants. When examining the importance of the latter group in classification, we notice that these are ranked in the last quartiles of the feature ranking for the Profiling model, except for gender information when predicting the informativeness of the review, where it appears in the second quartile. In contrast, demographic features are highly informative for the N-grams model, generally appearing in the first or second quartiles. Notably, gender is the most informative feature for all questions except Question 2, where it is the tenth in the ranking, preceded by various *n*-grams and the feature capturing reading habits. Interestingly, the user's

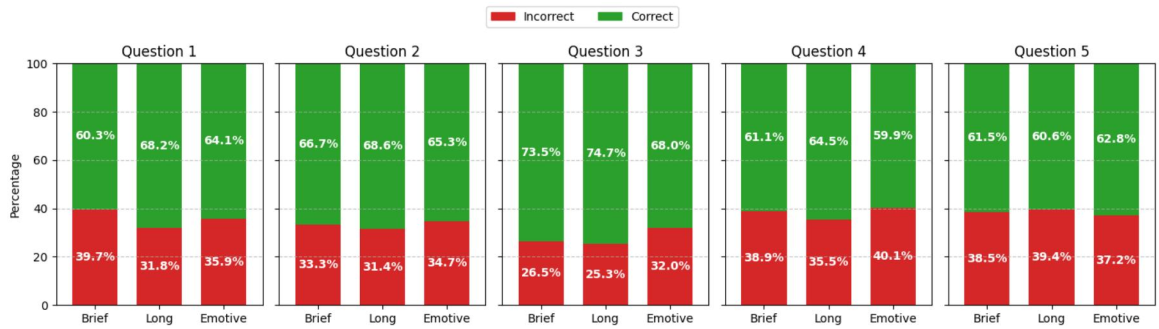


Figure 7. Distributions of correct and incorrect predictions of the Profiling model across the Brief, Long, and Emotive review groups.

generation is the least informative feature, consistently appearing in the fourth quartile for Questions 4 and 5. This aligns with the literature and our demographic analysis, which indicates that birth year is not a discriminative factor in assessing user preferences.

These findings suggest that although the Profiling and N-grams models yield the same classification accuracy, the Profiling model relies much less on demographic information, a type of information that is harder to collect than linguistic features from the reviews. This is an encouraging result, demonstrating that it's possible to reliably predict aspects of perception, including the level of enjoyment and willingness to read the reviewed book, using only the internal characteristics of the review itself, without needing to reconstruct the reader's personal profile.

7.1.2 Review style and prediction

As a final step in our analysis, we focused on the Profiling model, which offers the most promising real-world applications. Specifically, we examined whether the model's predictions are related to the writing style of the reviews. To investigate this, we analyzed the distribution of correct and incorrect predictions across brief, emotive, and long reviews. The results are presented in Fig. 7.

Comparatively, long reviews have the lowest percentage of prediction errors, while emotive reviews pose the greatest challenge for the model. Interestingly, this trend is reversed for Question 5, though the differences in misclassification rates across review types remain minimal, reinforcing the notion that predicting user enjoyment is inherently difficult in all scenarios. Additionally, for Question 1, the model predictions are less accurate for reviews composed of brief sentences. This is unsurprising, considering that scores assigned to evaluate the informativeness of brief reviews exhibited the lowest agreement in our dataset (see Table 3). We further observe that reviews with a high percentage of emotive lexicon are the most frequently misclassified for the three questions specifically related to writing style, i.e., Questions 2, 3, and 4. This aligns with our expectations, as the distinguishing features of emotive reviews are more closely tied to their lexicon rather than the morpho-syntactic and syntactic characteristics used as classification features by the Profiling model.

Overall, these findings highlight the impact of writing style on classification performance. While the Profiling model demonstrates robustness, its varying accuracy across different review types suggests that incorporating additional linguistic cues or structural elements could further enhance its reliability.

8. Discussion and conclusion

In this study, we examine a key aspect of Digital Social Reading practices: amateur book reviewing. Specifically, we introduce an innovative methodology that combines Computational Stylometry and Machine Learning to investigate what makes a book review compelling for readers, whose judgments were elicited through a survey. Our approach focuses in particular on identifying the writing characteristics of a corpus of Italian amateur reviews that most influence readers' perceptions of various aspects that contribute to making a review compelling, namely, informativeness, writing style, and general enjoyment, ultimately guiding readers toward reading the reviewed book. Our findings highlight the multifaceted relationship between readers' perception, the linguistic characteristics that shape a review's writing style, and demographic factors specific to different reader groups. We also show how these elements interact in the automatic prediction of review perception, offering valuable insights into reader engagement with book reviews.

In line with our initial research questions, this study explores various aspects of how the reading experience is communicated and received. Concerning the former, our analysis of reviewers' writing styles (RQ1) reveals that, beyond sentence length and word choice, linguistic features associated in the literature with syntactic structural complexity play a key role in distinguishing different review styles.

Our findings also provide insights into the dynamics of reader perception, particularly in relation to review writing style and reader demographics (RQ2). The results of the linear mixed-effects model used to assess the statistical significance of score variations across the five questions of the survey indicate that, despite the high degree of disagreement among raters, coherent patterns in how different aspects of amateur reviews are perceived do emerge. In addition, the test confirms that the surface-level characteristics used to categorize the reviews (i.e. sentence length and emotive lexicon) are significantly associated with differences in perception. Specifically, reviews in the Long category are perceived as more informative and more professional than those in the Brief category, which are instead associated with a more personal tone. Additional correlation analyses yield consistent results and further clarify how participants interpret review aspects. Taken together, these results suggest that participants, even with different personal interpretations of the questions eliciting review perception, rely on shared underlying intuitions that produce coherent patterns at the aggregated level. In particular, the data suggest that personal and professional tone are perceived not as independent traits, but rather as opposite ends of a single continuum. Thus, when subjects identify a review as having a personal style, they also coherently assign a low score to the question on the professional style. Furthermore, informativeness correlates positively with both dimensions of enjoyment (Questions 4 and 5), which themselves exhibit a very strong correlation, indicating that readers often associate informative reviews with greater overall appeal.

In line with previous research, our experiments investigating whether grouping survey responses by demographic traits and reading habits could reduce subjectivity in review perception did not yield strong results. The statistical tests showed that differences between groups were significant in only a few cases. These findings support the idea, also suggested in earlier studies, that spontaneous common interests or shared life experiences may be more influential than demographic traits in shaping amateur review perception.

The outcomes of our final research question (RQ3) mostly confirm the trends and findings previously discussed. Notably, our results show that predictive models struggled the most with predicting the perception of writing style enjoyment and global enjoyment, two aspects found to be highly subjective, as indicated by the high level of disagreement among survey participants. In contrast, the perception of the professional tone of the review emerged as the most predictable aspect, as it correlated with a more homogeneous set of linguistic characteristics used as features by the predictive models. Another revealing outcome concerns the comparative analysis of the two predictive models we developed: although the Profiling and N-grams models achieved comparable accuracy, the Profiling model relied significantly less on demographic information, suggesting that review perception can be effectively predicted using linguistic features alone. This finding highlights the feasibility of automatically assessing review engagement without requiring extensive reader profiling. Moreover, our results indicate that Long reviews have the lowest prediction errors, whereas Emotive reviews pose the greatest challenge for the models. This aligns with the idea that Long reviews tend to follow more predictable syntactic and lexical patterns, while Emotive reviews, characterized by expressive and subjective language, introduce greater variability.

In conclusion, our study contributes to a deeper understanding of Digital Social Reading, particularly the book reviewing process and its readers' perceptions. By focusing on the aspects that make a review compelling and influence readers' decisions to engage with the reviewed book, our findings may support future research on book recommendation systems. Specifically, we suggest that such systems could reliably predict a user's potential interest in a book based on the linguistic characteristics of its reviews, features that effectively model writing style while being far less invasive to collect than demographic or private user data.

Author contributions

Chiara Alzetta (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing), Felice Dell'Orletta (Conceptualization, Funding acquisition, Project administration, Supervision, Writing—review & editing), Alessio Miaschi (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing—original draft, Writing—review & editing), and Giulia Venturi (Conceptualization, Investigation, Methodology, Resources, Validation, Visualization, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Digital Scholarship in the Humanities* online.

Conflicts of interest. None declared.

Funding

Alessio Miaschi acknowledges the support of the Premio Buzzetti 2023 awarded by the “Associazione per l’Informatica Umanistica e la Cultura Digitale” (AIUCD).

Data Availability

The data underlying this article are partially available as online supplementary material and partially on reasonable request to the corresponding authors.

Notes

1. Book titles are reported as Supplementary Material SM1.
2. Note that we verified that the Brief and Long review groups contain a significantly lower amount of emotive lexicon, ensuring that they function as appropriate non-emotive counterparts to the Emotive group.
3. English translation of the review text: “I understand why some may feel lost among these pages, but for me, it was like a fantastic return home. The plot itself is rather predictable, but in this case, the journey is much more important than the destination. Recommended for all those nostalgic for the ’80s, especially if they are gamers, but I don’t think I can recommend it to everyone else. It’s a novel with prose that is smooth and enjoyable for various reasons, but the references to ’80s subcultures are so frequent that they can’t be ignored.”
4. Generation timeline (birth year interval): Baby Boomers (1946–1964), Generation X (1965–1980), Generation Y (1981–1996), Generation Z (1997–2012).
5. For the list of UD POS refer to https://universaldependencies.org/u/pos/index.html, of syntactic relations to https://universaldependencies.org/u/dep/index.html, and of verbal morphological tags https://universaldependencies.org/u/feat/index.html.
6. See Supplementary Material SM2 for the full set of linguistic characteristics with $P < .05$.
7. Questionnaire completion time was 00:19:45 (SD = ±00:09:13).
8. Results of the linear mixed-effects model are reported in Figures 1 and 2 (see online supplementary material for a colour version of these figures) of the Supplementary Material (SM4), also including the confidence intervals.
9. See Supplementary Material SM3 for the full set of linguistic characteristics that remain significant after FDR correction.
10. For agreement computing, we addressed the high prevalence of participants who read between four and eleven books per year by generating three random samples of similar size to the other reading habit groups and averaging their alpha scores.
11. Results of the linear mixed-effects model are reported in Figures 3–5 (see online supplementary material for a colour version of these figures) of the Supplementary Material (SM4), also including the confidence intervals.

References

Alex, A. B. (2019) ‘Reading Habits of Today’s Generation’, *International Journal on Global Business Management & Research*, 8: 69–74.

Alzetta, C. *et al.* (2024) ‘Tell Me How You Write and I’ll Tell You What You Read: A Study on the Writing Style of Book Reviews’, *Journal of Documentation*, 80: 180–202.

Biber, D. (1993) ‘Using Register-Diversified Corpora for General Language Studies’, *Computational Linguistics*, 19: 219–41. https://aclanthology.org/J93-2001/.

Boot, P., and Koolen, M. (2020) ‘Captivating, Splendid or Instructive? Assessing the Impact of Reading in Online Book Reviews’, *Scientific Study of Literature*, 10: 35–63.

Bronwen, T. (2020) *Literature and Social Media*. London: Routledge.

Brunato, D. *et al.* (2020) ‘Profiling-Ud: A Tool for Linguistic Profiling of Texts’, in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 7147–53, Marseille, France: European Language Resources Association.

Chaudhry, A. S., and Low, G. (2009) ‘Reading Preferences among Different Generations: A Study of Attitudes and Choices in Singapore’, *Singapore Journal of Library & Information Management*, 38: 27–48.

Chen, Y.-F. (2008) ‘Herd Behavior in Purchasing Books Online’, *Computers in Human Behavior*, 24: 1977–92.

Constantinides, E., and Holleschovsky, N. I. (2016) 'Impact of online product reviews on purchasing decisions', in *12th International Conference on Web Information Systems and Technologies, WEBIST 2016*, Rome, Italy, April 23–25, 2016, pp. 271–8. SCITEPRESS.

de Marneffe, M. C. *et al.* (2021) 'Universal Dependencies', *Computational Linguistics*, 47: 1–54. https://doi.org/10.1162/COLI_A_00402, https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies.

De Mauro, T., and Chiari, I. (2016) 'Il Nuovo Vocabolario di Base Della Lingua Italiana', *Internazionale*. https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana, accessed 3 March 2023.

Dimitrov, S. *et al.* (2021) 'Goodreads versus Amazon: The Effect of Decoupling Book Reviewing and Book Selling', *Proceedings of the International AAAI Conference on Web and Social Media*, 9: 602–5.

Frazier, L. (1985) *Syntactic Complexity. Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, 129–89. Cambridge: Cambridge University Press.

Gibson, E. (1998) 'Linguistic Complexity: Locality of Syntactic Dependencies', *Cognition*, 24: 1–76. https://doi.org/10.1016/s0010-0277(98)00034-1

Gildea, D., and Temperley, D. (2010) 'Do Grammars Minimize Dependency Length?', *Cognitive Science*, 34: 286–310.

Hajibayova, L. (2019) 'Investigation of Goodreads' Reviews: Kakutanied, Deceived or Simply Honest?', *Journal of Documentation*, 75: 612–26.

Hegel, A. (2018) *Ocial Reading in the Digital Age*. Los Angeles: UCLA.

Hu, N. *et al.* (2011) 'Manipulation in Digital Word-of-Mouth: A Reality Check for Book Reviews', *Decision Support Systems*, 50: 627–35.

Hu, Y. *et al.* (2022) 'Complexities associated with user-generated book reviews in digital libraries: temporal, cultural, and political case studies', in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pp 1–12.

Islam, M. A., and Muna, J. I. (2024) 'Do Generation z Students Read? reading Habits of the Students of Dhaka University', *Library Management*, 45: 597–612.

Istat - Istituto Nazionale di Statistica. (2023) Lettura di libri e fruizione delle biblioteche – Istat, periodo di riferimento: 2022. https://www.istat.it/comunicato-stampa/lettura-di-libri-e-fruizione-delle-biblioteche/

Koolen, M., Boot, P., and van Zundert, J. J. (2020) 'Online book reviews and the computational modelling of reading impact', in *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, pp. 1–21. CEUR.

Korfiatis, N., Rodriguez, D., and Sicilia, M.-A. (2008) 'The impact of readability on the usefulness of online product reviews: a case study on an online bookstore', in *Emerging Technologies and Information Systems for the Knowledge Society: First World Summit on the Knowledge Society, WSKS 2008, Athens, Greece, September 24–6, 2008. Proceedings 1*, pp. 423–32. Springer.

Korfiatis, N., García-Bariocanal, E., and Sánchez-Alonso, S. (2012) 'Evaluating Content Quality and Helpfulness of Online Product Reviews: The Interplay of Review Helpfulness vs. review Content', *Electronic Commerce Research and Applications*, 11: 205–17.

Kousha, K., Thelwall, M., and Abdoli, M. (2017) 'Goodreads Reviews to Assess the Wider Impacts of Books', *Journal of the Association for Information Science and Technology*, 68: 2004–16.

Kutzner, K., Schoormann, T., and Knackstedt, R. (2021) 'Exploring the Content Composition of Online Book Reviews', in *Informatik 2020*, pp. 1335–44. Bonn: Gesellschaft für Informatik.

Li, Z. *et al.* (2023) 'Bookgpt: A General Framework for Book Recommendation Empowered by Large Language Model', *Electronics*, 12: 4654.

Maity, S. K., Panigrahi, A., and Mukherjee, A. (2019) 'Analyzing Social Book Reading Behavior on Goodreads and How It Predicts Amazon Best Sellers', in M. Kaya and R. Alhajj (eds) *Influence and Behavior Analysis in Social Networks and Social Media, ASONAM 2018, Lecture Notes in Social Networks*, pp. 211–35. Cham: Springer.

Miranda, A. E., Hermansyah, H., and Mortini, A. V. (2023) 'Reading Habits in the Digital Age: Changes Trend and Behavior in Alpha Generation', *Esteem Journal of English Education Study Programme*, 6: 100–10.

Pianzola, F. (2025) *Digital Social Reading: Sharing Fiction in the Twenty-First Century*. Cambridge, USA: MIT Press.

Pianzola, F., Toccu, M., and Viviani, M. (2022) 'Readers' Engagement through Digital Social Reading on Twitter: The Twletteratura Case Study', *Library Hi Tech*, 40: 1305–21.

Prabowo, T. T. *et al.* (2025) 'Effect of Demographic Factors on Reading Habits: Study on Indonesian Secondary School Students', *Journal of Librarianship and Information Science*, 57: 705–18.

Rebora, S. *et al.* (2021) 'Digital Humanities and Digital Social Reading', *Digital Scholarship in the Humanities*, 36: ii230–50.

Sabri, N., and Weber, I. (2021) 'A Global Book Reading Dataset', *Data*, 6: 83.

Saraswat, M. (2022) ‘Leveraging Genre Classification with Rnn for Book Recommendation’, *International Journal of Information Technology*, 14:3751–6. https://doi.org/10.1007/s41870-022-00937-6

Sazzed, S. (2023) ‘What factors influence the popularity of user-generated text in the creative domain? a case study of book reviews’, in *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1957–62. IEEE.

Scofield, C. *et al.* (2022) ‘Book genre classification based on reviews of Portuguese-language literature’, in *International Conference on Computational Processing of the Portuguese Language*, pp. 188–97. Springer.

Spiteri, L. F., and Pecoskie, J. (2016) ‘Affective Taxonomies of the Reading Experience: Using User-Generated Reviews for Readers’ Advisory’, *Proceedings of the Association for Information Science and Technology*, 53: 1–9.

van Halteren, H. (2004) ‘Linguistic Profiling for Author Recognition and Verification’, in*Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 199–206. Barcelona, Spain.

Walsh, M., and Antoniak, M. (2021) ‘The Goodreads “Classics”: A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism’, *Journal of Cultural Analytics*, 4: 243–87. https://doi.org/10.22148/001c.22221

Wang, K., Liu, X., and Han, Y. (2019) ‘Exploring Goodreads Reviews for Book Impact Assessment’, *Journal of Informetrics*, 13: 874–86.

Wu, L. *et al.* (2017) ‘The Impact of Language Style on Consumers’ Reactions to Online Reviews’, *Tourism Management*, 59: 590–6. https://doi.org/10.1016/j.tourman.2016.09.006; https://www.sciencedirect.com/science/article/pii/S0261517716301662

Digital Scholarship in the Humanities, 2026, 00, 1–23
https://doi.org/10.1093/llc/fqag080
Full Paper