



HAL
open science

Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews

Chiara Alzetta, Felice Dell'Orletta, Alessio Miaschi, Elena Prat, Giulia Venturi

► To cite this version:

Chiara Alzetta, Felice Dell'Orletta, Alessio Miaschi, Elena Prat, Giulia Venturi. 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*, 2023, 10.1108/JD-04-2023-0073 . hal-04137185

HAL Id: hal-04137185

<https://hal.science/hal-04137185>

Submitted on 27 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL
open science

Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews

Chiara Alzetta, Felice Dell'Orletta, Alessio Miaschi, Elena Prat, Giulia Venturi

► To cite this version:

Chiara Alzetta, Felice Dell'Orletta, Alessio Miaschi, Elena Prat, Giulia Venturi. 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*, 2023, 10.1108/JD-04-2023-0073 . hal-04137185

HAL Id: hal-04137185

<https://hal.science/hal-04137185>

Submitted on 22 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**Tell Me How You Write and I'll Tell You What You Read:A
Study on the Writing Style of Book Reviews**

Journal:	<i>Journal of Documentation</i>
Manuscript ID	JD-04-2023-0073.R1
Manuscript Type:	Article
Keywords:	stylometric analysis, genre detection, book reviews, machine learning, reading platform, computational linguistics
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
main_revised.tex references_revised.bib	

SCHOLARONE™
Manuscripts

Tell Me How You Write and I'll Tell You What You Read: A Study on the Writing Style of Book Reviews

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

Anonymized for peer review

Abstract

Purpose - Our goal is to investigate variations in the writing style of book reviews published on different social reading platforms and referring to books of different genres, which enables acquiring insights **into communication strategies adopted by readers to share their reading experiences**.

Design/methodology/approach - We propose a corpus-based study focused on the analysis of *A Good Review*, a novel corpus of online book reviews written in Italian, posted on Amazon and Goodreads, and covering six **literary fiction** genres. We rely on stylometric analysis to explore the linguistic properties and lexicon of reviews and we conducted automatic classification experiments using multiple approaches and feature configurations to predict either the review's platform or the genre.

Findings - The analysis of user-generated reviews demonstrates that language is a quite variable dimension across reading platforms, but not as much across book genres. The classification experiments revealed that features modelling the syntactic structure of the sentence are reliable proxies for discerning Amazon and Goodreads reviews, whereas lexical information showed a higher predictive role for automatically discriminating the genre.

Originality - The high availability of cultural products makes information services necessary to help users navigate these resources and acquire information from unstructured data. This study contributes to a better understanding of the linguistic characteristics of user-generated book reviews, which can support the development of linguistically-informed recommendation services. Additionally, the authors release a novel corpus of online book reviews meant to support the reproducibility and advancements of the research.

Keywords: stylometric analysis, genre detection, book reviews, machine learning, reading platform, computational linguistics.

1 Introduction

The internet has revolutionized the way we read and write and led the way to the advent of Digital social reading (DSR), a series of practices devoted to the use of digital technologies and platforms to share with other people thoughts and impressions about the reading experience (Bronwen, 2020;

Pianzola, 2021). The popularity of such platforms has led to the creation of new social valences of reading (Nakamura, 2013) and, most importantly, of massive corpora of user-generated book reviews (see, e.g., Kousha et al., 2017; Sabri and Weber, 2021; Walsh and Antoniak, 2021). The insights that such reviews provide into readers' opinions are commonly exploited by publishers and authors for understanding the preferences of their readers (Aerts et al., 2017; Dimitrov et al., 2015; Maity et al., 2019; Thelwall, 2019; Wang et al., 2019). Among these studies, we notice that still little is known about the diverse communication strategies adopted by readers to share their reading experiences with others in terms of stylistic variations between reviews written across different platforms or referring to books belonging to different genres (Hajibayova, 2019; Reborá et al., 2021; Scofield et al., 2022), especially when it comes to book reviews written in Italian. However, the analysis of stylistic variations of reviewers' writing style can contribute useful perspectives to information behaviour research, especially in the context of Digital Social Reading practices. Not only it offers a novel perspective on the phenomena of information acquisition, use and sharing (Savolainen, 2020), but it can also offer new empirical data to gain insights into how readers communicate their diverse reading experiences across different types of books and platforms, thus allowing the investigation on the nature of the review enterprise as a vehicle of cultural production (Hajibayova, 2019). Additionally, stylistic analysis of book reviews holds value for library studies since it may foster the integration between digital libraries and new technologies for managing, organising and supporting the analysis of large amounts of unstructured information, which so far remains mostly an unexplored opportunity (Jeonghyun, 2021).

This study examines the linguistic properties and lexicon of Italian book reviews published on two leading platforms for DSR, Amazon Books¹ and Goodreads². The different purpose of these platforms makes them an interesting testbed to explore the diverse attitudes of online book reviewers: while Goodreads gathers a large community of amateur readers to exchange opinions and reading recommendations, Amazon has a marked commercial vocation and treats books mainly as a consumer good. Given their different nature, large-scale analyses of Goodreads users' behaviour were aimed at predicting the orientation of the book market (Maity et al., 2019; Wang et al., 2019), as well as to map reading preferences across various communities of users (Bourrier & Thelwall, 2020). Goodreads reviews have also proven to be an extremely valuable source to detect responses to reading (e.g. absorption) (Lendvai et al., 2020), and to analyze the linguistic style adopted by readers to describe their reading experiences (Driscoll & Rehberg Sedo, 2019; Nuttall & Harrison, 2020). Conversely, reviews posted on Amazon Books have mostly been investigated within marketing and buyers' behaviour studies, often relying on sentiment analysis (Chiavetta et al., 2016; Kaur & Singh, 2021; Srujan et al., 2018). Despite such widespread attention, there hasn't been a thorough comparison between them in terms of the communication strategies emerging from different writing styles in user-generated reviews. One of the few studies tackling this topic has suggested that, because of their inherent differences in terms of purpose (commercial/non-commercial), the two platforms elicit contrasting reading behaviours with regards to rating and review length (Dimitrov et al., 2015).

Our work contributes to this research area by introducing a novel corpus called *A Good Review*

¹<https://www.amazon.it/>

²<https://www.goodreads.com/>

1
2
3
4
5 (Amazon and GOODreads REVIEWS) which covers reviews of 300 books belonging to six literary
6 fiction genres and reviewed by users of Amazon and Goodreads. We expect book reviews published
7 on these two platforms to show multiple facets of readers' behaviour not only in terms of their
8 content but also accounting for the various communication strategies employed by readers. The
9 primary goal of the study is thus to investigate whether the stylistic variations between the reviews
10 are associated with different writing styles across platforms and genres.
11

12 Additionally, we explored whether stylistic properties of reviews allow us to automatically dis-
13 criminate the reviews' venue and to identify the genre of the reviewed book. To this aim, we
14 conducted two automatic classification experiments, Automatic Platform Classification (APC) and
15 Automatic Genre Classification (AGC), the task concerned with automatically detecting the genre
16 of textual documents (Kessler et al., 1997). While the first task is relatively straightforward, albeit
17 unexplored, due to the distinctive characteristics of the two platforms, the second task is more
18 challenging, particularly when using user-generated reviews as information sources.
19

20 Genre identification has been regarded as a key factor in library and information sciences (An-
21 dersen, 2008) for reducing inadequate results of search engines (Lim et al., 2004; Santini, 2007;
22 Santini, 2004), or for automatically enhancing documents with metadata (Lüscho & Tello, 2021).
23 In addition, AGC is employed in the publishing industry to develop book recommendation systems
24 (Alharthi et al., 2018) and in DSR activities such as the organization of books through digital
25 shelves and wish lists (Pianzola, 2021).
26

27 Despite the long-standing and widespread attention, the AGC task is extremely challenging due
28 to multiple factors. Previous works differ with respect to *i*) the notion of *genre*, which represents a
29 quite heterogeneous object of study (Biber & Conrad, 2009), *ii*) the typologies of text properties,
30 either token- (Crossley & Louwerse, 2007; HaCohen-Kerner et al., 2020; Mehler et al., 2010; Santini,
31 2004) or sentence-based characteristics (Cimino et al., 2017; Fang & Cao, 2010; Stamatatos et al.,
32 2001; Wan et al., 2019) or also accounting for the discourse structure of texts (Sun et al., 2021) *iii*)
33 the machine learning approach adopted (Worsham & Kalita, 2018), and *iv*) the source exploited to
34 classify the genre, either the book content (Rahul et al., 2021; Shamir, 2020; Worsham & Kalita,
35 2018), title or summary (Ozsarfati et al., 2019), descriptions on websites (Sobkowicz et al., 2018),
36 or cover design (Buczowski et al., 2018). Less attention has been paid to exploiting user-generated
37 reviews written by readers. To the best of our knowledge, the main exception is Saraswat, 2022,
38 which used Amazon book reviews for genre classification using recurrent neural networks, and the
39 study conducted by Scofield et al. (2022), which relied on multiple machine learning algorithms to
40 categorize 24 fiction genres starting from a collection of 325 Portuguese book reviews posted on
41 Goodreads.
42

43 The present study provides a comprehensive and detailed analysis of the lexical and stylistic
44 features of written communication on reading platforms and it contributes valuable evidence to the
45 field of Digital Social Reading activity, which has been lacking extensive empirical studies (Pianzola,
46 2021). The findings on the variations in the writing style of book reviews have implications for the
47 development of linguistically-informed recommender systems that offer personalized reading sugges-
48 tions: in an era where readers face a multitude of options, such systems have become increasingly
49 important (Alharthi et al., 2018). Information about the diverse writing styles of user-generated
50 book reviews could be leveraged, for example, to develop more effective book recommendation sys-
51
52
53
54
55
56

tems that target specific groups of readers with tailored communication strategies, as suggested by Ye et al., 2023. Overall, the study not only expands our understanding of written communication on reading platforms but also has practical applications in improving book recommendation systems. It bridges the gap in empirical research and provides valuable insights for both researchers and developers in the field.

2 The *A Good Review* Corpus

The present study relies on the *A Good Review* corpus (*Amazon and GOODreads REVIEWS*), a novel collection of book reviews acquired from Amazon and Goodreads across various *literary fiction* genres. The corpus is particularly suited for exploring whether the genre of a book impacts the writing style of its reviews and if reviews within the same genre have a distinctive writing style.

The creation of the *A Good Review* corpus starts by selecting the genres to be covered. Considering that the classification of books into *fiction* genres remains ambiguous, we relied on Goodreads' genre classification system based on 'shelves', i.e. user-curated collections complemented by brief descriptions specifying the type of books allowed in the list. Specifically, we identified six popular Goodreads 'shelves', corresponding to as many genres of narrative fiction: thriller, historical fiction, romance, science fiction, horror, and fantasy. We acquired the titles of the 50 most frequently reviewed books of each genre and used their ISBN number to unambiguously identify them on Amazon and Goodreads and to collect their reviews written in Italian. If a book appears on several shelves (e.g. *Frankenstein* classified both as horror and science fiction), it is assigned to the genre with which it is most frequently associated. Indeed, although Amazon users may not agree with the classification proposed on Goodreads platform, we expect that Amazon users are not likely to consistently disagree with it since the selection is based on the 50 most frequently assigned books in each genre. We obtained the reviews by extracting them from the Goodreads and Amazon websites using a web scraping technique. The reviews were gathered in June 2022.

Table 1: Dataset statistics for each genre and in total for each platform.

Genre	Amazon			Goodreads		
	Reviews	Sentences	Tokens	Reviews	Sentences	Tokens
Fantasy	8,608	31,229	567,472	8,316	75,402	1,701,735
Historical Fiction	4,455	16,050	296,361	5,196	43,486	1,037,555
Horror	3,958	16,677	329,801	6,219	51,815	1,205,617
Romance	6,885	28,970	527,996	7,855	76,992	1,707,373
Science Fiction	7,070	26,595	505,177	6,255	56,875	1,336,165
Thriller	5,952	21,699	383,449	4,847	34,484	765,947
Total	36,928	141,220	2,610,256	38,688	339,054	7,754,392

Table 1 describes the *A Good Review* corpus detailing the number of reviews collected from each platform and genre, as well as their length in terms of the number of sentences and tokens. We maintained the number of reviews balanced across platforms by collecting around 140 reviews for

each considered book. However, Goodreads reviews turned out to be longer than Amazon's, causing the Goodreads portion of the corpus to be larger. In fact, while Goodreads reviews feature around 200 tokens each, Amazon reviews show an average of around 70 tokens per review. The tendency to produce longer reviews on Goodreads is attested for all genres and represents a specific property of the data used in this study.

3 Approach

The study on variations in the writing style of book reviews is articulated in two main phases. In the first phase, we automatically acquired a set of stylistic properties (detailed in Section 3.1) from the reviews including (morpho-)syntactic and lexical properties. We analyzed the variation of these features across the review's venue (i.e. the platform), and the genre of the reviewed book. The goal was to determine if the platform has an impact on the writing style of the reviews and to compare the writing styles across different genres.

In the second phase, we conducted a series of classification experiments using multiple approaches and feature configurations described in Section 3.2. The experiments target two tasks: Automatic Platform Classification (APC) and Automatic Genre Classification (AGC). APC is devoted to predicting if a review was posted on either Amazon or Goodreads. In AGC, we aim at classifying the genre of the book being reviewed based on its review. AGC is addressed as a multi-class classification problem conducted separately for each platform by using only the reviews published on either Amazon or Goodreads. The different configurations of features used in both classification tasks are meant to investigate whether it is possible to predict the platform and the book's genre based on the review's writing style, as well as the effectiveness of various types of features in making this prediction.

3.1 Stylistic Properties

3.1.1 Linguistic Characteristics

The comparative analysis of the linguistic variation across the reviews has been inspired by research on linguistic profiling, a methodology originally developed for authorship recognition that detects and quantifies differences and similarities across texts representative of distinct language varieties relying on the distribution of a large number of linguistic features (van Halteren, 2004). In this study, we relied on the linguistic profiling methodology described in Brunato et al. (2020) and implemented in Profiling-UD, a web-based tool conceived to linguistically profile multilingual texts by relying on the Universal Dependencies (UD) formalism (De Marneffe et al., 2021), a *de facto* standard schema for morpho-syntactic annotation of corpora, based on the dependency syntactic representation paradigm. Figure 1 exemplifies the dependency tree representation based on the UD formalism of reviews acquired from the *A Good Review* corpus. As it can be noted, each word (token) of the two sentences is associated with a Part-Of-Speech (POS). Syntactic relations between words are represented as dependency links marked by a label indicating the type of relationship established between the two (e.g., `nsubj` for nominal subjects, `det` for determiners).

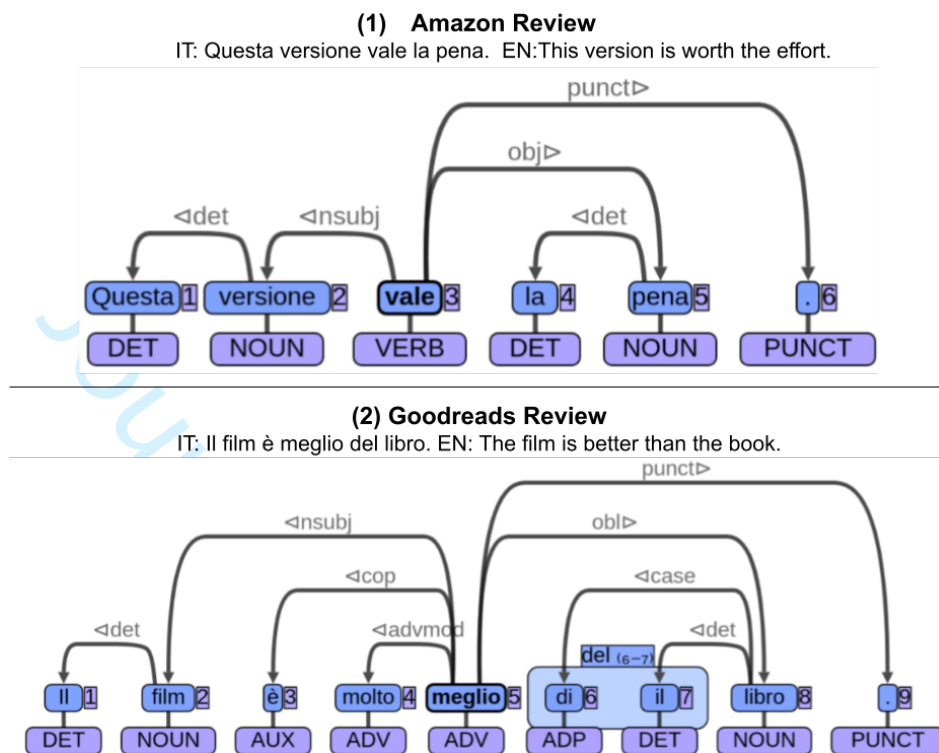


Figure 1: Amazon (1) and Goodreads (2) reviews linguistically annotated according to the UD schema.

Profiling-UD computes a large set of linguistic features derived from raw, morpho-syntactic, and syntactic levels of UD annotations. In this study, we relied on the 150 most frequent features occurring in the *A Good Review* corpus to prevent data sparsity issues.

As shown in Table 2, the features can be grouped according to nine main aspects. They range from quite simple aspects related to raw text properties (i.e. sentence and word length), to the distribution of UD Parts-Of-Speech and of inflectional properties specific in particular to verbal predicates (i.e. mood, tense, person).

More complex features concern the global and local syntactic structure of the UD dependency tree. For instance, the average length of dependency links, a feature occurring in the *Tree Structure* group, refers to a local property of a sentence and it is computed as the average number of words linearly occurring between a syntactic head and its dependent (excluding punctuation). Accordingly, in Figure 1, the average dependency length of review (1) is 1.25, and 1.86 in (2). The maximum link length complements the latter feature and refers to the longest link in the dependency tree. In (1), the longest link is the two-token long link connecting the pair [*pena* ‘effort’, *vale* ‘worth’]³, while in (2) we have two three-token long links: [*film* ‘movie’, *meglio* ‘better’] and [*libro* ‘book’, *meglio* ‘better’]. The maximum tree depth, on the other hand, represents a global property of the sentence structure, and it is computed as the path (in terms of occurring dependency links) from the root of

³The syntactic head is always the second element of the pair.

Table 2: Morpho-syntactic characteristics acquired from reviews.

Linguistic Feature	Label
Raw Text Properties (<i>RawText</i>)	
Document length	tokens, sentences
Avg sentence length	sent_length
Avg word length	char_per_tok
Vocabulary Richness (<i>Vocabulary</i>)	
Type/Token Ratio for words and lemmas (first 100 and 200 tokens)	ttr_F (100/200), ttr_L (100/200)
Lexical density	lexical_density
New Basic Italian Vocabulary for words and lemmas	NBIV
Fundamental/High usage/High availability words of NBIV for words and lemmas	in_FO_*, in_AU_*, in_AD_*
Morphosyntactic information (<i>POS</i>)	
Distribution of UD POS	POS label
Inflectional morphology (<i>VerbInflection</i>)	
Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Verbal Predicate Structure (<i>VerbPredicate</i>)	
Distribution of verbal heads and verbal roots	verbal_head, verbal_root_perc
Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
Global and Local Parsed Tree Structures (<i>TreeStructure</i>)	
Avg depth of syntactic trees	parse_depth
Avg length of dependency links and of the longest link	avg_links_len, max_links_len
Avg length of prepositional chains and distribution by depth	avg_prep_chain_len, prep_dist_1, prep_dist_*
Avg clause length	avg_token_per_clause
Order of elements (<i>Order</i>)	
Relative order of subject and object	subj_pre, subj_post, obj_pre, obj_post
Syntactic Relations (<i>SyntacticDep</i>)	
Distribution of dependency relations	dep label
Use of Subordination (<i>Subord</i>)	
Distribution of subordinate clauses	subord_prop_dist
Avg length of subordination chains and distribution by depth	avg_subord_chain_len, subord_dist_1
Relative order of subordinate clauses	subord_pre, subord_post

the tree to the furthest leaf node. In (1), it corresponds to the two intermediate dependency links that are crossed in the path going from the root of the sentence *vale* ‘worth’ to each of the most

distant leaf nodes, represented by the words *Questa* ‘This’, *la* ‘the’.

Among the set of considered features, we also include the quantitative distribution of a few vocabulary characteristics (*Vocabulary* group). This group covers a varied set of properties, such as the Type/Token Ratio (TTR), which refers to the ratio between the number of lexical types and the number of tokens within a text computed for both the first 100 and 200 tokens in the review, and the lexical density of a text, calculated as the percentage of content words in the review. Additionally, this group accounts for the distribution of word forms and lemmas belonging to the New Basic Italian Vocabulary (NBIV) (De Mauro & Chiari, 2016), also articulated into the usage classification classes of ‘fundamental’ (very frequent), ‘high usage’ (frequent) and ‘high availability’ words (relatively lower frequency words referring to everyday objects or actions and thus well known to speakers).

3.1.2 Lexical Variations

To complement the quantitative lexical analysis introduced above, we performed a qualitative analysis of the reviews’ lexicon devoted to identifying the main topics addressed by the readers, as well as the main semantic fields associated with reading experiences. This analysis is based on the frequency lists of words acquired from the linguistically annotated corpus of Amazon and Goodreads reviews. To reduce data sparsity, frequency lists are based on words base forms (lemmas) having at least 5 occurrences in the corpus and grouped by POS. In particular, the analysis focuses on words annotated as verbs (VERB), nouns (NOUN), and adjectives (ADJ). These are among the most relevant morpho-syntactic categories conveying the informational content of sentences in a document.

3.2 Models

The classification experiments rely on multiple machine learning (ML) algorithms that vary with respect to their architecture and features used for training. Specifically, as detailed below, we exploited a suite of classifiers based on Linear Support Vector Machines (LinearSVM) and one of the most prominent Neural Language Models (NLMs), i.e. BERT (Devlin et al., 2019). These models were trained and tested by adopting a 5-fold cross-validation approach, i.e. models are trained iteratively on $\frac{4}{5}$ of the reviews and use the remaining fifth as a test set. This ensured that the model was trained using a representative sample of the dataset at each iteration. The train and test sets always contained reviews of different books to ensure that those used for training are not seen during the test phase, thus increasing the complexity of the classification tasks. Models’ performance on both platform and genre classification tasks is evaluated based on Accuracy, Precision, Recall, and F-score.

3.2.1 Linear Support Vector Machine

LinearSVM is a machine learning discriminative algorithm that, given a set of training examples each marked as belonging to a class (e.g. Amazon and Goodreads, or one of the six genres), builds a model that assigns new unseen examples to one of the available classes. Specifically, an SVM model maps training examples to points in space in order to maximize the width of the gap between

1
2
3
4
5 classes. New examples are then mapped into that same space and assigned to one of the available
6 classes.

7 We chose this model over the other commonly used Machine Learning algorithms since it repre-
8 sents a reliable trade-off between the interpretability and accuracy of the results. In fact, similarly
9 to other models, it allows the identification of the most significant features that contribute to pre-
10 dicting the class of a document. However, it is particularly capable of handling high-dimensional
11 feature spaces and finding accurate and robust decision boundaries that can generalize well to new,
12 unseen documents.

13
14 For the purposes of this study, we define two LinearSVM models, referred to as Profiling and
15 Ngrams models. The former is linguistically informed since it takes the set of linguistic charac-
16 teristics extracted using the Profiling-UD tool as input features. The latter exploits only lexical
17 information since it uses as input feature a contiguous sequence of n words acquired from the reviews
18 (i.e. n-grams, with n equal to 1, 2, and 3).

21 22 3.2.2 Neural Language Model

23
24 Neural network models of language, i.e. Neural Language Models (NLMs), have recently become
25 the most commonly used algorithms for the resolution of NLP tasks. In particular, introducing the
26 Transformer model (Vaswani et al., 2017) allowed these algorithms to perform astonishingly well
27 across several tasks. For APC and AGC, we relied on one of the most prominent NLM based on
28 the Transformer architecture, i.e. BERT (Devlin et al., 2019). BERT (Bidirectional Encoder Rep-
29 resentations from Transformers) is an Encoder model trained to approximate the Masked Language
30 Modeling function, i.e. randomly masking part of the input tokens in a sentence and asking the
31 model to predict them, and fine-tuned for the resolution of a target task. For the purpose of the
32 present work, we relied on an Italian pre-trained version of the model (12 layers, 768 hidden units),
33 i.e. the one developed by the MDZ Digital Library Team and available through the Huggingface's
34 Transformers library (Wolf et al., 2020)⁴. The model was trained using the Italian Wikipedia and
35 the Italian portion of the OPUS corpus (Tiedemann & Nygaard, 2004), a multilingual collection of
36 translated open source documents available on the Internet.

37 38 39 40 41 3.2.3 LinearSVM + NLM

42 We combined the previous models into a classifier based on LinearSVM and trained using the
43 internal representations of the BERT model fine-tuned either on APC or AGC tasks. We refer to
44 this model as SVM (BERT). SVM (BERT+Profiling) is an additional LinearSVM model trained
45 using both the fine-tuned representations produced by BERT and Profiling-UD features. The BERT
46 representations used as input features of the SVM model result from averaging the embeddings of
47 all the tokens in each review.

48
49
50 ⁴<https://huggingface.co/dbmdz/bert-base-italian-xxl-cased>

3.2.4 Baselines

We compared the performance of the above models against multiple baselines. For what concerns APC, we relied on two baseline models: i) Majority class, which assigns the most frequent class as the output label for the review, and ii) Sent-length, a LinearSVM taking as input a single feature capturing the length of the review in terms of tokens. For AGC, we relied on the Sent-length model described above and on a random uniform classifier, i.e. a model that uniformly generates random predictions for each possible class (namely, all genres have equal prediction probability).

4 Analysis of Reviews' Style

The investigation of the writing style of the *A Good Review* corpus started with the analysis of the (morpho-)syntactic properties of reviews, discussed in Section 4.1. As a second study (see Section 4.2), we explored the lexicon used by the readers in terms of semantic domains adopted to describe their reading experience. In both investigations, we followed a two-level analysis, devoted on the one hand to highlighting differences and similarities across the two platforms, and on the other hand to comparing the style of reviews of books belonging to different genres. We refer to the former as 'platform-level analysis' and to the latter as 'genre-level analysis'.

4.1 Linguistic Analysis

4.1.1 Platform-level analysis

As a first step, we verified whether a statistical analysis reveals significant differences between the monitored linguistic characteristics when computed over the set of Amazon and Goodreads reviews. To this aim, we relied on the Mann-Whitney U rank test for independent samples, which revealed that all characteristics vary significantly ($p < 0.05$). This result suggests a quite different writing style across platforms. To assess which characteristics show major differences, we computed their rank-biserial correlation score r (Wendt, 1972), which ranges between +1 (when the distribution values of a given characteristic occurring in Goodreads reviews are higher than those in Amazon) and -1 (when the values are higher in Amazon reviews). The higher the r score, the greater the distribution difference between the platforms. Table 3 shows characteristics with $|r| \geq 0.2$ and their mean and standard deviation in Amazon and Goodreads reviews.⁵

Among the features showing higher $|r|$, we find the three raw text properties. As suggested in Section 2, it appears that users are more likely to write a long review on Goodreads. Accordingly, we count 5.4% (around 2,000) of reviews whose length corresponds to 1 or 2 tokens on Amazon, while they are only 1% (less than 400) on Goodreads. Additionally, readers tend to write longer sentences when they write on Goodreads (*sent_length*).

In the set of the most varying features, we also find characteristics modeling aspects related to the local syntactic structure of a sentence (*TreeStructure*). In particular, Goodreads sentences contain a higher number of embedded complement chains governed by a nominal head (*n_prep_chains*), and

⁵See Appendix for the full set of Profiling-UD features.

Table 3: Platform-level feature analysis: for each characteristic whose $|r| \geq 0.2$, we report the mean and standard deviation values computed on Amazon and Goodreads reviews. Features in each group are ordered by decreasing rank-biserial correlation value (r).

Group	Feature	Amazon	Goodreads	r
		Mean (stdev)	Mean (stdev)	
RawText	tokens	70.69 (± 128.56)	200.43 (± 286.13)	(+).416
	sentences	3.82 (± 5.24)	8.76 (± 11.70)	(+).385
	sent_length	16.38 (± 10.87)	20.80 (± 11.22)	(+).270
Vocab.	ttr_F (100)	0.10 (± 0.26)	0.35 (± 0.37)	(+).333
	ttr_L (100)	0.09 (± 0.22)	0.30 (± 0.33)	(+).332
	lexical_density	0.56 (± 0.14)	0.51 (± 0.09)	(-).235
	NBIV	0.87 (± 0.14)	0.84 (± 0.12)	(-).230
	ttr_F (200)	0.04 (± 0.15)	0.18 (± 0.29)	(+).218
	ttr_L (200)	0.03 (± 0.13)	0.15 (± 0.25)	(+).218
	POS	PROPN	2.10 (± 7.37)	2.72 (± 4.52)
ADJ		12.56 (± 16.28)	8.09 (± 8.05)	(-).220
NUM		0.55 (± 1.85)	0.99 (± 3.37)	(+).217
Verb	aux_3 st prs-plr	7.11 (± 18.23)	10.70 (± 17.56)	(+).215
Inflection	verbs_3 st prs-plr	7.80 (± 19.45)	11.48 (± 18.20)	(+).213
Verb	verbal_heads	2.01 (± 1.60)	2.51 (± 1.54)	(+).237
Predicate	verb_edges_dist_5	4.51 (± 10.94)	6.38 (± 9.81)	(+).234
	avg_verb_edges	2.21 (± 1.06)	2.53 (± 0.86)	(+).214
Tree Structure	n_prep_chains	2.63 (± 5.23)	7.32 (± 11.09)	(+).366
	max_links_len	11.88 (± 9.53)	18.97 (± 13.66)	(+).351
	avg_links_len	2.25 (± 0.67)	2.52 (± 0.52)	(+).282
	avg_max_links_len	7.29 (± 5.19)	8.99 (± 4.92)	(+).252
	avg_prep_chain_len	0.78 (± 0.59)	0.94 (± 0.48)	(+).214
	avg_max_depth	3.67 (± 1.85)	4.23 (± 1.74)	(+).206
Syntactic Dep	root	12.06 (± 17.81)	7.28 (± 8.77)	(-).270
	det:poss	0.61 (± 1.64)	0.84 (± 1.41)	(+).234
	expl	0.78 (± 1.58)	1.06 (± 1.42)	(+).225
	iobj	0.59 (± 1.48)	0.72 (± 1.31)	(+).206
	ccomp	0.67 (± 1.42)	0.84 (± 1.32)	(+).201
Subord	nummod	0.45 (± 1.37)	0.74 (± 2.45)	(+).200
	avg_sub_chain_len	0.95 (± 0.61)	1.11 (± 0.52)	(+).229
	subord_dist_2	11.99 (± 23.48)	15.54 (± 19.86)	(+).212

longer syntactic dependency relations (*avg_links_len*, *max_links_len*). Note that all the above represent aspects of a more complex and articulated writing style (Frazier, 1985; Gibson, 1998; Gildea & Temperley, 2010).

Amazon and Goodreads book reviews strongly differ also with respect to their lexical richness. Namely, Goodreads readers tend to use a more diverse lexicon, corresponding to lower values of

1
2
3
4
5 words belonging to the NBIV (*NBIV*) and higher values of Type/token ratio both at the level
6 of forms and lemma computed for the first 100 words of the document (ttr_F_100 , ttr_L_100).
7 Our intuition is that on a domain-specific platform such as Goodreads users possibly describe
8 several aspects of their reading experiences. **In contrast**, reviews on a more commercial platform
9 such as Amazon are restricted to a few semantic domains. This is what we investigated in the
10 qualitative analysis of our study (Section 4.2). The relatively high rank-biserial correlation score of
11 the distribution of proper nouns (*PROPN*), more frequent on Goodreads, could be similarly related
12 to the diverse content of the book reviews. A qualitative investigation revealed that Goodreads
13 reviews tend to contain more names of book characters and of places where the story is set, while
14 Amazon readers tend to refer more frequently to publishing companies, actors and directors involved
15 in film adaptations, and e-reader devices.
16
17
18

19 4.1.2 Genre-level analysis

20
21 To investigate cross-genre variations, we analysed the distribution of linguistic characteristics in
22 the reviews of books belonging to the six considered **fiction** genres. We release these distributions
23 as supplementary materials. Below, we discuss characteristics showing significant variations across
24 genres as determined by the Mann-Whitney U statistical test. Table 4 shows the percentage of
25 characteristics of a group that vary significantly ($p < 0.05$) for a genre compared to all the others.
26 The higher the number of varying characteristics, the more different the review style for that genre.
27
28

29 Note that even though a total high number of characteristics varies across genres (i.e. about
30 50-60%, see *All* row), the six genres only share a very small subset of characteristics that vary
31 significantly for all of them (i.e. 19 on Goodreads and 4 on Amazon, out of 150). This seems to
32 indicate that Amazon's reviews have a more homogenous writing style across genres compared to
33 Goodreads' since they have a lower number of the same characteristics that vary significantly for
34 all genres. Additionally, their distribution values are quite similar, as confirmed by the low r scores
35 (see Appendix) which are not worth discussing in more depth. We will limit ourselves to point out
36 that, among the characteristics showing the highest r scores, we find many features falling in the
37 *Vocabulary* group, suggesting that variations in reviews' style may mostly lie in different use of the
38 lexicon.
39
40

41 However, if we focus on the amount and type of statistically varying characteristics rather than
42 on their values, we notice that on both platforms each genre differs from the others with respect to a
43 different group of linguistic characteristics. In most cases, the writing style varies significantly con-
44 sidering two groups of characteristics: raw text and the distribution of morpho-syntactic categories
45 (*POS*). The latter is also among the characteristics showing a greater variation for most genres,
46 based on their r score. As already noted, these groups contain linguistic phenomena which can pos-
47 sibly affect the distribution of other characteristics of a sentence. For example, longer sentences are
48 possibly characterized by deeper syntactic trees or longer dependency relations; sentences containing
49 a higher percentage of verbs or conjunctions may feature more subordinate constructions. Horror
50 and Fantasy books on Goodreads represent exceptions: they differ significantly in the syntactic
51 structure of their sentences and the order of subjects and objects with respect to the main verb
52 (*Order*). Thriller books on Amazon, on the other hand, vary mainly for vocabulary characteristics.
53
54
55
56
57
58
59
60

Table 4: For each platform and genre, we report (in percentage) the number of linguistic characteristics within groups that vary significantly ($p < 0.05$) compared to all the other genres. Results are ordered based on the values of the *All* row, i.e. the percentage obtained considering the whole set of significant characteristics.

Features Group	Amazon					
	Fantasy	Horror	Romance	Science Fiction	Thriller	Historical Fiction
Raw text	100	100	50	100	25	25
POS	64.71	70.59	76.47	58.82	58.82	70.59
VerbInflection	60.53	47.37	47.37	50	52.63	42.11
VerbPredicate	80	50	70	60	40	40
TreeStructure	72.73	90.91	54.55	81.82	27.27	54.55
Order	75	50	25	100	50	50
SyntacticDep	62.79	62.79	65.12	44.19	48.84	48.84
Subord	70	60	30	60	50	60
Vocabulary	53.85	76.92	69.23	30.77	92.31	53.85
All	65.33	62.67	58	54	52	50
Features Group	Goodreads					
	Thriller	Romance	Historical Fiction	Horror	Fantasy	Science Fiction
Raw text	100	75	50	50	50	100
POS	82.35	82.35	76.47	64.71	58.82	76.47
VerbInflection	81.58	55.26	36.84	50	42.11	39.47
VerbPredicate	60	40	60	40	50	–
TreeStructure	81.82	81.82	72.73	72.73	63.64	90.91
Order	50	50	50	50	75	50
SyntacticDep	46.51	76.74	72.09	48.84	48.84	48.84
Subord	50	60	50	40	40	20
Vocabulary	84.62	69.23	53.85	69.23	53.85	61.54
All	68	67.33	58.67	53.33	50	50

Interestingly, we notice that readers of the same genre tend to adopt a different writing style based on the platform where their review is published. For example, the reviews of Thriller books written on Amazon differ for a higher percentage of characteristics (68%) than those written on Goodreads (52%). They also vary mostly with respect to different groups: raw text (100%) for Goodreads and vocabulary (92.31%) for Amazon. In fact, Thriller reviews posted on Goodreads are on average shorter than the other reviews and contain shorter sentences and words. Conversely, reviews posted on Amazon feature a higher percentage of NBIV words belonging in particular to the Fundamental repository and are characterized by a lower lexical variability (in terms of TTR), compared to the reviews of the other genres.

Table 5: For each genre, the frequency distribution (expressed in percentage) of words according to their grammatical category. Lexical overlap and Spearman correlation coefficients ($p < 0.05$) are computed for each genre between word lists acquired from Amazon and Goodreads.

POS	Genre	Frequency Distribution		Lexical Overlap	Spearman Correlation
		Amazon	Goodreads		
Nouns	Fantasy	17.10	15.53	42.11	.770
	Historical Fiction	17.67	16.76	40.84	.814
	Horror	17.01	16.69	39.94	.805
	Romance	16.09	15.07	40.80	.823
	Science Fiction	17.41	16.68	44.51	.821
	Thriller	17.04	16.48	50.10	.834
	All	16.97	15.99	45.57	.837
Adjectives	Fantasy	7.76	6.37	42.78	.768
	Historical Fiction	7.99	6.63	38.73	.820
	Horror	7.28	6.69	37.42	.811
	Romance	7.58	6.30	41.28	.837
	Science Fiction	7.90	6.72	43.02	.829
	Thriller	7.87	6.60	50.58	.839
	All	7.73	6.50	46.92	.834
Verbs	Fantasy	10.25	10.56	45.29	.826
	Historical Fiction	10.37	10.21	43.67	.852
	Horror	10.25	10.25	44.77	.856
	Romance	11.01	10.86	43.09	.887
	Science Fiction	10.27	10.23	48.40	.861
	Thriller	10.54	10.53	55.65	.861
	All	10.48	10.50	46.86	.901

4.2 Lexical Analysis

4.2.1 Platform-level analysis

The qualitative investigation of the reviews' lexicon focuses on the nouns, verbs and adjectives having at least five occurrences in the corpus. Their distributions across genres and overall for each platform are reported in Table 5.

As can be noted, the frequency distribution values of the three considered POS are quite similar across platforms and genres despite the reviews' different lengths. The lower distribution of nouns and adjectives on Goodreads reviews is simply due to the greater presence of function words introducing, e.g., subordinates and embedded complement chains (see Section 4.1), which lowers the relative amount of content words on this platform.

To determine if this similarity also extends to the use of the lexicon, we used two metrics, presented in Table 5. *Lexical overlap* measures the rate of words used in both Amazon and Goodreads reviews. This is computed – for each POS and overall – as a ratio of the lemmas appearing at

1
2
3
4
5 least once in both platforms and the total amount of distinct lemmas in the *A Good Review* corpus.
6 *Spearman's rank correlation coefficient*, on the other hand, is meant to determine if words sharing
7 the same grammatical category occur with a similar relative frequency in both platforms, which
8 would indicate that readers refer to the same topics.
9

10 Table 5 reveals that the reviews posted on the two platforms share about half of their lexicon (see
11 lexical overlap values above 45% for all POS). If we look at this value in light of the high Spearman
12 correlation scores, we can conclude that readers rely on a similar lexicon on both platforms to
13 convey their opinions.
14

15 In order to further analyse this shared lexicon and the recurring topics that it might convey,
16 we manually categorized the adjectives, verbs and nouns appearing within Amazon and Goodreads
17 reviews into one of the eight semantic classes listed in Table 6. We defined the classes on the basis of
18 previous research on recurring topics, reading-related metaphors and signs of emotional engagement
19 in online reading reviews (Dimitrov et al., 2015; Nuttall & Harrison, 2020; Pianzola et al., 2020).
20 The result of this process is an annotated lexicon comprising 1,074 lemmas, each assigned to a
21 semantic class which we distribute for future research.
22

23 Based on the frequency distribution values, we noticed that some classes are equally represented
24 in the reviews of Amazon and Goodreads. Consider in particular the case of the *Other media*
25 and *Reading-related* classes. The former covers terms related to the transposition of a book into
26 another audiovisual medium and to artistic performances, also not directly related to reading (e.g.
27 *cantare* 'to sing'). The *Reading-related* class includes terms referring to narrative elements and
28 the act of reading and it is also the most represented class of annotated lexicon (around 55% on
29 both platforms). Interestingly, the names of the six genres appear in this class among the most
30 frequent terms. This not only confirms our assumption that users of both Amazon and Goodreads
31 share a similar understanding of the main **fiction** genres, but also shows that they are very much
32 concerned with assigning books to a genre. In addition, we notice that reviewers on Goodreads and
33 Amazon also tend to attach great significance to the pace of their reading, as suggested by the high
34 frequency of terms such as *incalzante* 'fast-paced', as well as to the level of likelihood of the story
35 (e.e., *credibile* 'credible', *inverosimile* 'unlikely').
36
37
38

39 Conversely, some classes are more represented within Amazon reviews, even though their distri-
40 bution percentages remain quite low on both platforms. These are the *Advice-related* class, which
41 covers terms explicitly related to sharing advice and ratings, the *Commercial* class, which includes
42 terms referring to the book's price and shipping conditions, and the *Book's materiality* class, con-
43 taining terms referring to material properties of the books. The frequency gap is especially marked
44 within the *Commercial* class (5.26% on Amazon, 1.79% on Goodreads). This shows that Amazon
45 users are prone to evaluate books also as generic consumer goods. This assumption is strengthened
46 by the higher frequency, on Amazon, of terms related to the visual and tactile features of the book,
47 which range from aesthetic elements (*illustrazione* 'illustration') to types of binding. Moreover, we
48 noticed that some terms falling into this category - e.g. *imballato* 'packed' and *graffiato* 'scratched' -
49 only appear in the Amazon dataset. Finally, the slightly higher frequency of the *Advice-related* class
50 suggests that Amazon users tend to turn to this platform more for giving and receiving advice on
51 whether to buy and read a book, rather than for commenting on their reading experience. Overall,
52 these three classes point towards a more spread concern with practical aspects of books and book
53
54
55
56
57
58
59
60

Table 6: Distribution of lexical classes across Amazon and Goodreads platforms. Examples cover words annotated at the POS level as adjectives (A), nouns (N) and verbs (V).

Class	Frequency Distribution %		Examples
	Amazon	Goodreads	
Reading-related	55.41	55.59	A: <i>classico</i> ‘classic’, <i>scorrevole</i> ‘smooth’; N: <i>libro</i> ‘book’, <i>storia</i> ‘story’; V: <i>leggere</i> ‘to read’, <i>narrare</i> ‘to recount’
Affect-related	21.07	26.79	A: <i>interessante</i> ‘interesting’, <i>avvincente</i> ‘engaging’; N: <i>emozione</i> ‘emotion’, <i>aspettativa</i> ‘expectation’; V: <i>deludere</i> ‘to disappoint’, <i>piacere</i> ‘to like’
Advice-related	4.69	3.27	A: <i>consigliato</i> ‘recommended’; N: <i>recensione</i> ‘review’, <i>giudizio</i> ‘rating’; V: <i>suggerire</i> ‘to suggest’
Book’s materiality	5.43	3.21	A: <i>rigido</i> ‘hardcover’, <i>flessibile</i> ‘paperback’; N: <i>pubblicazione</i> ‘publication’, <i>illustrazione</i> ‘illustration’; V: <i>rilegare</i> ‘to bind’, <i>stampare</i> ‘to print’
Commercial	5.26	1.79	A: <i>gratuito</i> ‘for free’, <i>caro</i> ‘pricy’; N: <i>acquisto</i> ‘purchase’, <i>corriere</i> ‘courier’; V: <i>prenotare</i> ‘to book’, <i>regalare</i> ‘to give’
Other media	3.08	2.83	A: <i>cinematografico</i> ‘cinematographic’, <i>teatrale</i> ‘theatrical’; N: <i>film</i> ‘movie’, <i>smartphone</i> ; V: <i>cantare</i> ‘to sing’, <i>dipingere</i> ‘to paint’
Travel-related	3.76	4.52	A: n/a N: <i>mondo</i> ‘world’, <i>viaggio</i> ‘trip’; V: <i>partire</i> ‘to leave’, <i>viaggiare</i> ‘to travel’
Food-related	1.30	1.99	A: <i>amaro</i> ‘bitter’, <i>piccante</i> ‘spicy’; N: <i>gusto</i> ‘taste’, <i>cena</i> ‘dinner’; V: <i>assaporare</i> ‘to taste’, <i>divorare</i> ‘to devour’

circulation among Amazon readers.

In contrast, Goodreads users seem to be more concerned with the exploration of their own subjective reading responses. This is suggested by the higher frequency of the *Affect-related* class on Goodreads compared to Amazon, as shown in Table 6. This class encompasses the expressions of feelings, ideas and opinions driven (more or less explicitly) by an affect. Drawing on core theories in the field of affect theory and emotion studies (Scherer, 2005; Tomkins, 1984), we included terms pertaining to basic emotions (e.g. *noia* ‘boredom’), to the bodily reactions associated with them (e.g. *piangere* ‘to cry’), to aesthetic emotions (e.g. *ammirare* ‘to be in awe’), as well as terms generically referring to affective phenomena, including moods (e.g. *allegro* ‘cheerful’), affect dispositions (e.g. *geloso* ‘jealous’) and preferences (e.g. *appassionato* ‘passionate’).

Finally, Table 6 also reports two semantic classes - *Food-related* and *Travel-related* - which include terms possibly associated with two recurring metaphors used to discuss the intensity of readers’ engagement with a book: metaphors related to eating (e.g. I have devoured this book) and to travel or transportation (e.g. I was transported by the story). Whilst the data might

Table 7: Distribution of the annotated lexical classes according to genres on Amazon (A) and Goodreads (G).

		Advice	Commer- cial	Food	Book Mater.	Other media	Reading	Affect	Travel
Fantasy	A	4.67	7.17	0.89	9.28	3.02	53.41	16.72	4.83
	G	3.31	1.63	1.39	3.65	2.57	58.03	23.52	5.90
Hist-Fi	A	4.72	3.39	1.29	3.88	2.12	59.75	21.29	3.55
	G	3.16	1.56	1.61	2.98	1.92	59.28	25.10	4.38
Horror	A	4.25	4.05	1.31	5.59	3.65	58.48	18.78	3.87
	G	3.04	1.72	1.53	3.48	2.98	59.81	23.44	4.01
Sci-Fi	A	4.47	4.37	1.06	4.40	4.63	57.85	18.86	4.36
	G	3.23	1.83	1.57	3.15	3.98	57.73	22.58	5.93
Thriller	A	5.18	3.96	1.09	4.24	2.91	59.81	20.08	2.73
	G	3.62	1.81	1.81	3.68	2.80	61.85	21.08	3.34
Romance	A	4.78	4.60	1.84	3.68	2.15	52.69	27.50	2.76
	G	3.65	1.82	2.17	2.97	1.90	52.61	31.59	3.30

point towards a more frequent use of metaphorical reading-related language on Goodreads, the low percentages of the annotated lexicon call for caution when evaluating these results.

4.2.2 Genre-level analysis

Having assessed how the terms annotated with the semantic classes are distributed on the two platforms, we proceeded to examine their distribution across the six genres.

The results are reported in Table 7. Overall, the class distributions are consistent with the percentages displayed in Table 6, with some noticeable exceptions which seem to be mainly driven by the recurring narrative elements of selected genres. For instance, we noticed that the class *Other media* is the most frequent one within the Science Fiction reviews posted on both platforms: this might suggest that Science Fiction readers tend the most to compare the book to the existing film adaptation and/or to employ a lexicon related to other media in order to discuss the plot and the characters. This is not surprising, given that Science Fiction books often touch upon the topic of other media, in particular technology.

A closer observation of the affect-related lexicon offers some hints as to how a certain genre might inform the readers' perceptions and responses. For Horror and Thriller reviews, we have noticed a high frequency of terms relating to the fulfilment or disappointment of the reader's expectations, such as 'deludere' to *disappoint* and 'sorpriendente' *surprising*. Readers of Thriller and Horror want to be surprised and occasionally end up being disappointed. As expected, reviews of horror also feature a high frequency of terms relating to fear and bewilderment, such as 'paura' *fear* and 'inquietante' *unsettling*.

In contrast, reviews of Romance and Historical Fiction feature a stronger presence of terms relating to being moved - such as 'commovente' *moving* - as well as generic terms relating to the sphere of emotions and love, such as 'amare' to *love* and 'sentimento' *feeling*.

Table 8: APC Results.

Model	Accuracy	Precision	Recall	F-score
Majority class	0.31	0.39	0.61	0.48
Sent-length	0.53	0.51	0.61	0.56
Profiling	0.64	0.63	0.66	0.64
Ngrams	0.59	0.55	0.63	0.59
BERT	0.82	0.86	0.76	0.81
SVM (BERT)	0.86	0.85	0.89	0.87
SVM (BERT+Profiling)	0.86	0.85	0.89	0.87

Lastly, it is interesting to notice that terms belonging to the *Commercial* and the *Book's materiality* classes are particularly frequent in reviews of Fantasy books posted by Amazon users, possibly due to the higher reviewing activity of fantasy readers on this platform.

5 Classification Experiments

5.1 Automatic Platform Classification

Table 8 reports the results of the APC task in terms of *i*) Precision, measured as the ratio between the number of correct predictions for a platform and the total amount of reviews assigned by a model to that platform, *ii*) Recall, i.e. the ratio between the number of correct predictions for a platform and the number of reviews actually belonging to each platform, and *iii*) F-score which combines the previous two metrics since it represents the harmonic mean of the model's Precision and Recall.

As a general remark, we observe that all models outperform the two baselines Majority class and Sent-length. However, models that exploit BERT representations achieve the highest results, especially when used as training features for the SVM model. Interestingly, although one may assume that reviews are easily distinguished based on their lexical properties, the linguistic features of Profiling-UD drive higher results than n-grams (0.64 vs. 0.59 F-score). This suggests that morpho-syntactic and syntactic properties can be used as reliable proxies for discerning reviews written by Amazon or Goodreads users. However, adding linguistic features to the SVM model trained with BERT fine-tuned representations does not provide any performance improvement. The result seems to suggest that BERT internal representations already implicitly encode the (morpho-)syntactic information explicitly conveyed by the set of linguistic features considered, as previously shown by Miaschi et al., 2020; Rogers et al., 2020.

5.2 Automatic Genre Classification

Table 9 reports the results obtained by the classification models for the AGC task we performed AGC separately for Amazon and Goodreads reviews. For each genre, we computed precision and

Table 9: AGC results: precision (P), recall (R), F-score (F) and accuracy (A) on Amazon (A) and Goodreads (G) reviews.

Genre	M	Random Uniform		Sent Length		Profiling		Ngrams		BERT		SVM (BERT)		SVM (BERT+ Profiling)	
		A	G	A	G	A	G	A	G	A	G	A	G	A	G
Hor.	P	.11	.17	.12	0	.15	.22	.47	.55	.7	.62	.58	.8	.57	.8
	R	.18	.18	.06	0	.15	.17	.45	.51	.5	.59	.58	.77	.58	.76
	F	.14	.17	.08	0	.15	.19	.46	.53	.58	.6	.58	.78	.57	.78
Hist-Fi	P	.09	.13	0	0	.18	.22	.33	.46	.52	.56	.5	.73	.5	.73
	R	.16	.15	0	0	.19	.26	.46	.45	.37	.52	.5	.77	.5	.76
	F	.11	.14	0	0	.18	.24	.38	.45	.43	.54	.5	.75	.5	.74
Sci-Fi	P	.25	.16	.05	.08	.23	.23	.65	.55	.61	.72	.65	.78	.65	.78
	R	.17	.16	0	.01	.18	.19	.54	.51	.53	.63	.57	.77	.57	.78
	F	.2	.16	.01	.01	.2	.21	.59	.53	.57	.67	.6	.77	.61	.78
Thrill.	P	.21	.12	0	.10	.24	.2	.57	.53	.59	.61	.63	.72	.62	.72
	R	.18	.17	0	.17	.24	.24	.49	.53	.54	.58	.56	.76	.56	.76
	F	.19	.14	0	.13	.24	.22	.53	.53	.56	.6	.59	.74	.59	.74
Rom.	P	.26	.2	.22	.18	.26	.28	.58	.56	.54	.61	.63	.79	.62	.79
	R	.17	.17	.03	.22	.27	.34	.5	.58	.51	.68	.6	.79	.59	.77
	F	.21	.18	.06	.20	.27	.31	.54	.57	.52	.64	.61	.79	.6	.78
Fant.	P	.12	.22	.11	.19	.32	.28	.32	.56	.54	.66	.6	.81	.6	.81
	R	.16	.17	.83	.47	.36	.24	.53	.59	.79	.68	.73	.78	.71	.79
	F	.14	.19	.19	.27	.34	.26	.4	.58	.64	.67	.66	.8	.65	.8
All	A	.17	.17	.11	.17	.25	.24	.5	.54	.57	.63	.6	.78	.6	.77
	P	.17	.17	.08	.09	.23	.24	.49	.53	.58	.63	.6	.77	.59	.77
	R	.16	.17	.06	.10	.23	.24	.48	.53	.55	.62	.59	.77	.59	.77
	F	.16	.17	.06	.10	.23	.24	.48	.53	.55	.62	.59	.77	.59	.77

recall as above, whereas, for the scores obtained on the full corpus (All), we report their overall unweighted mean.

As can be noted, the tested models outperform the baselines in all the configurations, but, unlike APC, the SVM trained using the linguistic characteristics (Profiling), shows the lowest performance overall and for each genre. In fact, as observed in Section 4.1, the distribution values of the (morpho-)syntactic characteristics are quite similar across the six genres on both platforms. On the other hand, higher results are obtained when relying on models that exploit lexical information, such as Ngrams and BERT. As already noted for the APC task, the best-performing model turned out to be SVM (BERT), which points to the lexical and semantic nature of the AGC task. Interestingly, adding the set of linguistic features extracted by Profiling-UD to the SVM trained with BERT representations did not improve the results. Similarly to what was observed above, this could suggest that Transformer-based models already encode within their representations information about the morpho-syntactic and syntactic structure of a document and, therefore, explicitly adding

1
2
3
4
5 such information does not enhance the discriminatory ability of the model for this task.

6 When looking at the differences between the two platforms, we can observe that the scores
7 are generally higher on Goodreads reviews. This seems in line with what we observed in Sec-
8 tion 4.1, where we highlighted that the readers' writing style on Goodreads is more heterogeneous
9 across different genres, thus probably making it easier for the classifier to distinguish between their
10 corresponding reviews. This can be noticed in particular for the models that exploit BERT repre-
11 sentations as input features for the SVM model. **In contrast**, few differences can be observed when
12 the (morpho-)syntactic characteristics are used. A quite peculiar case is represented by Fantasy,
13 which is the only genre whose reviews have been classified with higher accuracy when written on
14 Amazon. Notably, it is also the best-classified genre by the three models that exploit BERT's rep-
15 resentations, regardless of the platform. Further investigations of this result are discussed in the
16 following Section.
17
18
19

20 21 **5.2.1 Error Analysis**

22 Figure 2 complements the above results by reporting the confusion matrices with the percentage of
23 the predictions made by all models. The analysis shows:
24

- 25
26 • which genres are more confusing. Regardless of the platform, we observe that, as the model
27 performance improves, the erroneous predictions tend to be quite similarly distributed over
28 all possible genres. In line with the overall higher classification results, this is particularly
29 the case of Goodreads matrices obtained by the analysis of the SVM (BERT) and SVM
30 (BERT+Profiling) models. Conversely, the matrices of the Profiling and Ngrams models are
31 the most sparse. Consider for example the Goodreads Historical Fiction reviews: they have
32 been confused most of the time with the Romance ones, i.e. 20% of times by the Profiling
33 model and 17% by the Ngrams one. This is possibly due to the fact that books belonging to
34 these genres often have plots at the crossroads between the two genres;
35
36
- 37
38 • which are the most wrongly classified genres. It resulted that Romance is the most confus-
39 ing genre for Goodreads reviews in particular for Profiling-UD, while for Amazon the most
40 misleading genre is Fantasy, which is the most predicted class, even more frequently than the
41 correct one. Note that it is the most confusing genre also for the other classification models
42 including BERT. Our intuition is that the high frequency of Fantasy and Romance reviews
43 in *A Good Review* could explain these errors. In fact, according to the statistics about the
44 internal composition of the corpus (cf. Table 1), they represent the most numerous classes in
45 terms of number of tokens. This may suggest that the length of the review may impact the
46 accuracy of the AGC task.
47
48

49 To investigate this hypothesis further, we explored the models' performance variation with
50 respect to the review length (in tokens). Namely, we ordered Amazon and Goodreads reviews
51 based on their increasing length and split the rankings into ten portions (bins) of equal size: each
52 bin comprises 3,691 reviews for Amazon and 3,867 for Goodreads. Bins are aimed at covering
53
54
55
56

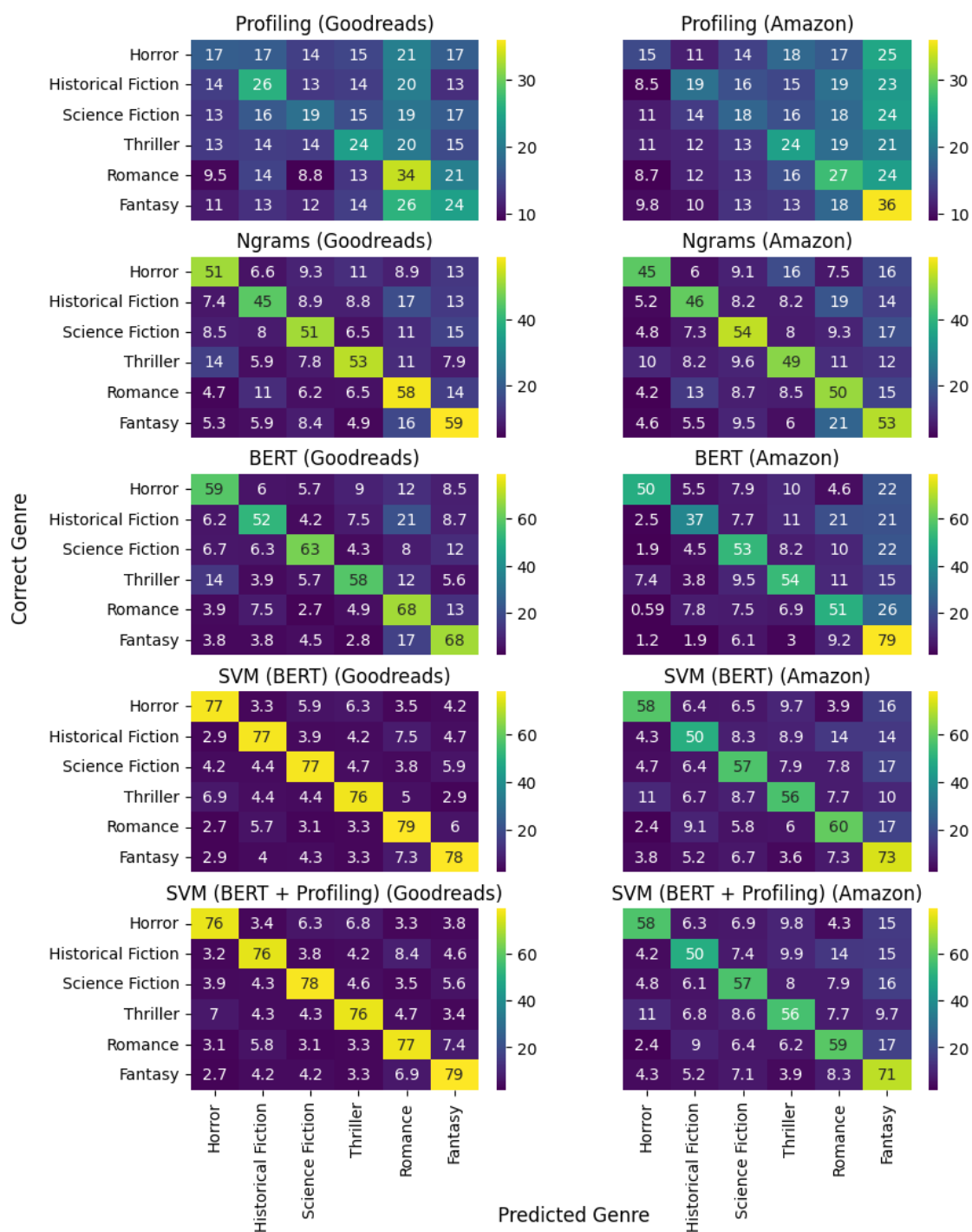


Figure 2: Confusion matrices of the AGC task for all models: cells report the percentage of reviews automatically assigned to a genre by each classification model (column) with respect to their actual genre (row). The diagonal refers to the percentage of correct predictions.

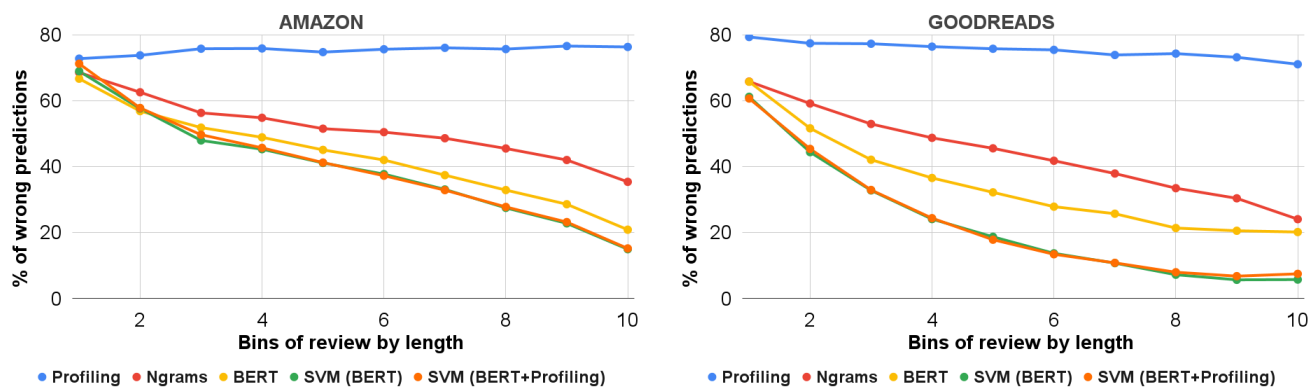


Figure 3: For the AGC task, the percentage of incorrect predictions of each model at different intervals of review length.

different ranges of review lengths, from very small (bin 1 = 7.77 tokens on average) to very long reviews (bin 2 = 855.83 tokens on average)⁶.

Figure 3 shows the percentage of incorrect predictions made by each model for each bin on the two platforms. As can be observed, the models' trends are quite similar also across platforms. The Profiling model is the one showing the most different performance, as it makes a similar amount of classification errors regardless of the reviews' length. Conversely, the other models reduce the percentage of errors as the length of the reviews increases. Interestingly, although Ngrams achieved lower results than BERT, the two models made approximately the same amount of errors when classifying long reviews (> 491 tokens). Possibly, the Ngrams model keeps learning as the review gets longer, whereas BERT is restricted to the sequence length it was trained with (i.e. 512 subwords maximum). As seen in Table 9, SVM (BERT) and SVM (BERT+Profiling) show nearly identical performance.

6 Conclusions

The work introduced *A Good Review*, a new corpus of Italian book reviews posted on Amazon and Goodreads, which allowed investigating multi-faceted writing-style variations across the two platforms and across **fiction** genres.

Specifically, the stylistic analysis shows that Goodreads' users tend to write longer reviews characterized by a more complex and articulated writing style. This is in contrast to a previous comparative analysis conducted by Dimitrov et al., 2015, and further investigations could determine whether this difference is language-dependent. In addition, it revealed that on Amazon, the readers' writing style is more homogeneous across genres compared to Goodreads, and readers of the same genre tend to adopt a different style based on the platform on which their review is published. Regarding the use of the lexicon, we highlight a greater use of commercial-related terms and terms

⁶The distribution values of the (morpho-)syntactic characteristics across the 10 bins are reported as supplementary materials in Appendix.

related to the material properties of the book on Amazon and a different distribution of terms referring to readers' affective responses across genres. This seems to support the hypothesis that certain genres are more likely to reach specific areas of the emotional sphere of readers, or at least to lead them to express their feelings (Kim et al., 2017). This analysis was based on a lexicon manually annotated with semantic classes, which we distribute to support researchers interested in analysing book reviews or in developing software programs or applications for text classification, sentiment analysis, or book recommendation.

The two classification experiments enabled us to assess the impact of various stylistic properties extracted from reviews on the prediction of the venue and genre of the books reviewed. The results showed that lexical information, in particular those contained in BERT's deep semantic representations, are more predictive than properties modeling the syntactic structure of the sentence. This emerges clearly in genre classification, which points to the lexical and semantic nature of the task. Furthermore, review length turned out to be a crucial factor for the task as longer reviews were more accurately predicted than shorter ones. However, in line with the high variations observed in the stylistic analysis, the linguistic features of Profiling-UD proved to be reliable proxies for discerning Amazon and Goodreads reviews.

The study offers several directions for future research. Firstly, the approach adopted can foster multilingual studies since the set of linguistic characteristics we considered is based on the UD formalism, which guarantees the comparative encoding of language phenomena across different languages (Nivre, 2015). Additionally, the *A Good Review* corpus and the annotated lexicon could be expanded with novel genres and books. Furthermore, the promising results of the classification experiments suggest that user-generated reviews can be exploited to develop linguistically-informed book recommendation systems providing personalized reading recommendations based on a user's writing style. **Finally, in a broader perspective, we believe that this study can contribute to the understanding of the complex and multifaceted phenomenon of Digital Social Reading, taking as an innovative starting point the user-generated book reviews. It is precisely the reviews, understood as cultural products, and the communication strategies adopted by readers, which become themselves writers, that provide a privileged observation point on the practice of reading as a collective sharing activity of individual experiences.**

References

- Aerts, G., Smits, T., & Verlegh, P. W. (2017). How online consumer reviews are influenced by the language and valence of prior reviews: A construal level perspective. *Computers in Human Behavior*, 75, 855–864.
- Alharthi, H., Inkpen, D., & Szpakowicz, S. (2018). A survey of book recommender systems. *Journal of Intelligent Information Systems*, 51, 139–160.
- Andersen, J. (2008). The concept of genre in information studies. *Annual review of information science and technology*, 42(1), 339.
- Biber, D., & Conrad, S. (2009). *Genre, register, style*. Cambridge University Press.
- Bourrier, K., & Thelwall, M. (2020). The social lives of books: Reading victorian literature on goodreads. *J. of Cultural Analytics*, 5(1), 12049.

- 1
2
3
4
5 Bronwen, T. (2020). *Literature and social media*. Routledge.
- 6 Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-UD: A
7 tool for linguistic profiling of texts. *Proc. of LREC*, 7147–7153.
- 8 Buczkowski, P., Sobkowicz, A., & Kozłowski, M. (2018). Deep learning approaches towards book
9 covers classification. *Proc. of ICPRAM* (pp. 309–316).
- 10 Chiavetta, F., Bosco, G. L., & Pilato, G. (2016). A lexicon-based approach for sentiment classifica-
11 tion of amazon books reviews in italian language. *WEBIST (2)*, 2016, 159–170.
- 12 Cimino, A., Wieling, M., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2017). Identifying predic-
13 tive features for textual genre classification: The key role of syntax. In R. Basili, M. Nissim,
14 & G. Satta (Eds.), *Proc. of CLiC-it*. CEUR.
- 15 Crossley, S. A., & Louwse, M. (2007). Multi-dimensional register classification using bigrams. *Int.*
16 *J. of Corpus Linguistics*, 453–478.
- 17 De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Com-*
18 *putational linguistics*, 47(2), 255–308.
- 19 De Mauro, T., & Chiari, I. (2016). Il nuovo vocabolario di base della lingua italiana. *Internazionale*
20 *[accessed on 06/04/2023]*. <https://intern.az/1w1U>
- 21 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional
22 transformers for language understanding. *Proc. of NAACL, vol.1*, 4171–4186.
- 23 Dimitrov, S., Zamal, F., Piper, A., & Ruths, D. (2015). Goodreads versus amazon: The effect of
24 decoupling book reviewing and book selling. *Proc. of Int. AAAI ICWSM*, 9(1), 602–605.
- 25 Driscoll, B., & Rehberg Sedo, D. (2019). Faraway, so close: Seeing the intimacy in goodreads reviews.
26 *Qualitative Inquiry*, 25(3), 248–259.
- 27 Fang, A. C., & Cao, J. (2010). Enhanced genre classification through linguistically fine-grained pos
28 tag. *Proc. of PACLIC*, 223–232.
- 29 Frazier, L. (1985). Syntactic complexity. *Natural Language Parsing: Psychological, Computational,*
30 *and Theoretical Perspectives*, Cambridge: Cambridge University Press, 129–189.
- 31 Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 24(11),
32 1–76.
- 33 Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*,
34 34(2), 286–310.
- 35 HaCohen-Kerner, Y., Daniel, M., & Yigal, Y. (2020). The influence of pre-processing on text clas-
36 sification using a bag-of-words representation. *PloS one*, 15(5).
- 37 Hajibayova, L. (2019). Investigation of goodreads' reviews: Kakutanied, deceived or simply honest?
38 *J. of Documentation*, 75(3), 612–626.
- 39 Jeonghyun, K. (2021). Determining research data services maturity: The role of library leadership
40 and stakeholder involvement. *Library & Information Science Research*, 101092.
- 41 Kaur, K., & Singh, T. (2021). Impact of online consumer reviews on amazon books sales: Empirical
42 evidence from india. *J. of Theoretical and Applied Electronic Commerce Research*, 16(7),
43 2793–2807.
- 44 Kessler, B., Numberg, G., & Schütze, H. (1997). Automatic detection of text genre. *Proc. of EAACL*,
45 32–38.
- 46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

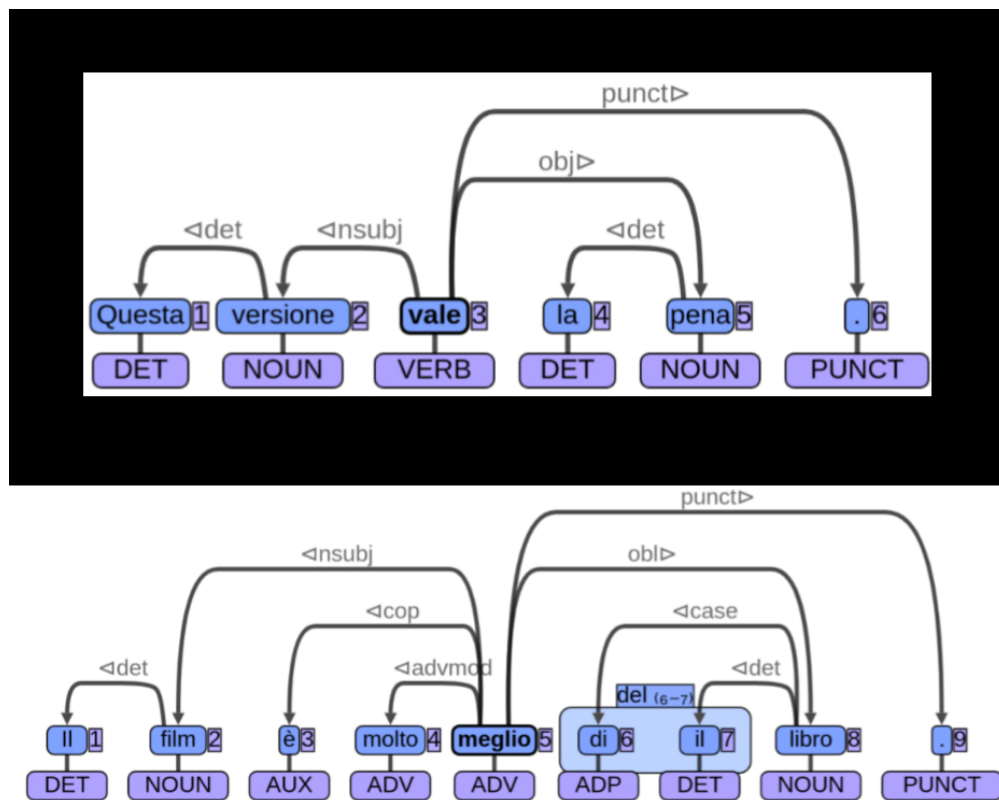
- 1
2
3
4
5 Kim, E., Padó, S., & Klinger, R. (2017). Prototypical emotion developments in literary genres. *Proc. of SIGHUM Workshop*, 17–26.
- 6
7 Kousha, K., Thelwall, M., & Abdoli, M. (2017). Goodreads reviews to assess the wider impacts of
8 books. *JASIST*, 68(8), 2004–2016.
- 9
10 Lendvai, P., Darányi, S., Geng, C., Kuijper, M., de Lacalle, O. L., Mensonides, J.-C., Reborá, S.,
11 & Reichel, U. D. (2020). Detection of reading absorption in user-generated book reviews:
12 Resources creation and evaluation. *LREC 2020-12th Conference on Language Resources and*
13 *Evaluation*, 4835–4841.
- 14
15 Lim, C., Kong, L., & Gil, K. (2004). Multiple sets of features for automatic genre classification of
16 web documents. *Information processing and management*, 41, 1263–1276.
- 17
18 Lüscho, A., & Tello, J. C. (2021). Towards genre classification in the library catalog. *Proceedings*
19 *of the Conference on Digital Curation Technologies (Qurator 2021)*.
- 20
21 Maity, S. K., Panigrahi, A., & Mukherjee, A. (2019). Analyzing social book reading behavior on
22 goodreads and how it predicts amazon best sellers. *Influence and Behavior Analysis in Social*
23 *Networks and Social Media*, 211–235.
- 24
25 Mehler, A., Sharoff, S., & Santini, M. (2010). *Genres on the web: Computational models and em-*
26 *pirical studies* (Vol. 42). Springer Science & Business Media.
- 27
28 Miaschi, A., Brunato, D., Dell’Orletta, F., & Venturi, G. (2020). Linguistic profiling of a neural
29 language model. *Proc. of COLING*, 745–756.
- 30
31 Nakamura, L. (2013). “Words with friends”: Socially networked reading on goodreads. *Pmla*, 128(1),
32 238–243.
- 33
34 Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh
35 (Ed.), *Computational linguistics and intelligent text processing* (pp. 3–16). Springer.
- 36
37 Nuttall, L., & Harrison, C. (2020). Wolfing down the twilight series: Metaphors for reading in online
38 reviews. *Contemporary media stylistics*, 35–60.
- 39
40 Ozsarfatí, E., Sahin, E., Saul, C. J., & Yilmaz, A. (2019). Book genre classification based on titles
41 with comparative machine learning algorithms. *Proc. of ICCCS*, 14–20.
- 42
43 Pianzola, F. (2021). *Digital social reading: Sharing fiction in the 21st century*. MIT Press.
- 44
45 Pianzola, F., Reborá, S., & Lauer, G. (2020). Wattpad as a resource for literary studies. quantitative
46 and qualitative examples of the importance of digital social reading and readers’ comments
47 in the margins. *PloS one*, 15(1).
- 48
49 Rahul, Ayush, Agarwal, D., & Vijay, D. (2021). Genre classification using character networks. *Proc.*
50 *of ICICCS*, 216–222.
- 51
52 Reborá, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M. M.,
53 Lauer, G., Lendvai, P., Messerli, T. C., & Sorrentino, P. (2021). Digital humanities and
54 digital social reading. *Digital Scholarship in the Humanities*, 36(Supplement₂), ii230–ii250.
- 55
56 Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about
57 how BERT works. *TACL*, 8, 842–866.
- 58
59 Sabri, N., & Weber, I. (2021). A global book reading dataset. *Data*, 6(8), 83.
- 60
Santini, M. (2007). Enhanced genre classification through linguistically fine-grained pos tag. *Proc.*
of PACLIC, 223–232.

- 1
2
3
4
5 Santini, M. (2004). Identification of genres on the web: A multi-faceted approach. *Language Learning and Technology*, 1–8.
- 6
7 Saraswat, M. (2022). Leveraging genre classification with rnn for book recommendation. *Int. J. of*
8 *Information Technology*, 1–6.
- 9
10 Savolainen, R. (2020). Sharing information through book reviews in blogs: The viewpoint of rosen-
11 blatt’s reader-response theory. *Journal of Documentation*, 76(2), 440–461.
- 12
13 Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science infor-*
14 *mation*, 44(4), 695–729.
- 15
16 Scofield, C., Silva, M. O., de Melo-Gomes, L., & Moro, M. M. (2022). Book genre classification
based on reviews of portuguese-language literature. *Proc. of PROPOR*, 188–197.
- 17
18 Shamir, L. (2020). UDAT: Compound quantitative analysis of text using machine learning. *Digital*
Scholarship in the Humanities, 36(1), 187–208.
- 19
20 Sobkowicz, A., Kozłowski, M., & Buczkowski, P. (2018). Reading book by the cover - Book genre de-
21 tection using short descriptions. In e. a. Gruca A. (Ed.), *Man-machine interactions* (pp. 439–
22 448). Springer.
- 23
24 Srujan, K., Nikhil, S., Raghav Rao, H., Karthik, K., Harish, B., & Keerthi Kumar, H. (2018).
Classification of amazon book reviews based on sentiment analysis. *Information systems*
25 *design and intelligent applications* (pp. 401–411). Springer.
- 26
27 Stamatatos, E., Nikos, F., & George, K. (2001). Automatic text categorization in terms of genre
28 and author. *Computational Linguistics*, 26, 471–495.
- 29
30 Sun, K., Wang, R., & Xiong, W. (2021). Investigating genre distinctions through discourse distance
and discourse network. *Corpus Linguistics and Linguistic Theory*, 17(3), 599–624.
- 31
32 Thelwall, M. (2019). Reader and author gender and genre in goodreads. *Journal of Librarianship*
and Information Science, 51(2), 403–430.
- 33
34 Tiedemann, J., & Nygaard, L. (2004). The OPUS corpus - parallel and free. *Proc. of LREC*.
- 35
36 Tomkins, S. S. (1984). Affect theory. *Approaches to emotion*, 163(163-195), 31–65.
- 37
38 van Halteren, H. (2004). Linguistic profiling for author recognition and verification. *Proc. of ACL*,
200–207.
- 39
40 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polo-
sukhin, I. (2017). Attention is all you need. In I. Guyon & alii (Eds.), *Advances in neural*
information processing systems. Curran Associates, Inc.
- 41
42 Walsh, M., & Antoniak, M. (2021). The goodreads “classics”: A computational study of readers,
amazon, and crowdsourced amateur criticism. *J. of Cultural Analytics*, 4, 243–287.
- 43
44
45 Wan, C., Yuling, W., Yaoze, L., Jinchao, J., & Guozhong, F. (2019). Composite feature extraction
and selection for text classification. *IEEE*, 7, 35208–35219.
- 46
47 Wang, K., Liu, X., & Han, Y. (2019). Exploring goodreads reviews for book impact assessment. *J.*
of Informetrics, 13(3), 874–886.
- 48
49
50 Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank-biserial
coefficient of correlation based on the statistic. *European J. of Social Psychology*.
- 51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5 Worsham, J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in
6 literature. *Proc. of COLING*, 1963–1973.
7
8 Ye, J., Xiong, H., Guo, J., & Meng, X. (2023). A group recommender system for books based on
9 fine-grained classification of comments. *The Electronic Library*, 1–21. <https://doi.org/https://doi.org/10.1108/EL-11-2022-0252>
10
11
12
13

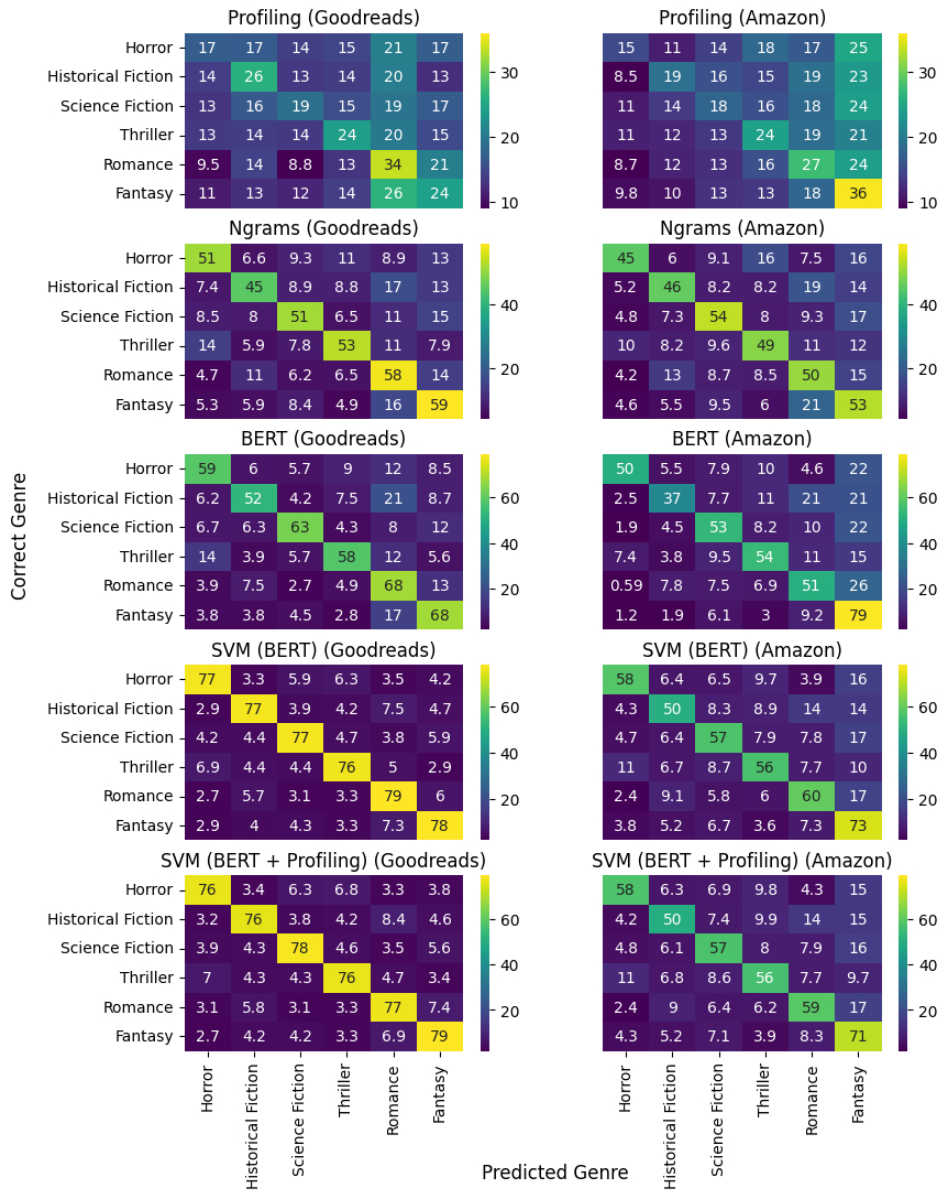
14 Appendix

15
16
17 The supplementary files are available at this link
18 https://drive.google.com/drive/folders/1Og4MWKZbh44WIDljqT97_bvd8xz1_2E6.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



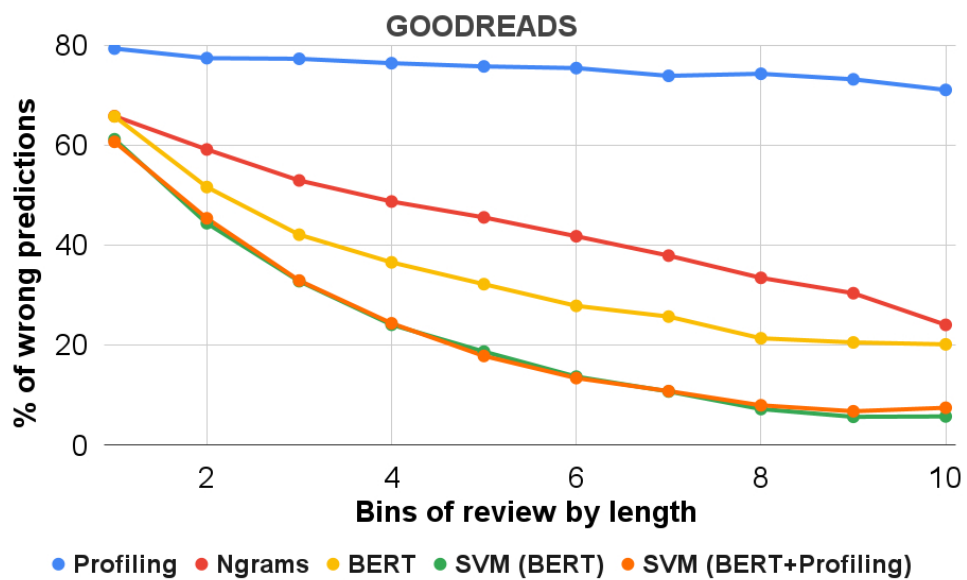
Amazon (1) and Goodreads (2) reviews linguistically annotated according to the UD schema.

327x261mm (72 x 72 DPI)



Confusion matrices of the AGC task for all models: cells report the percentage of reviews automatically assigned to a genre by each classification model (column) with respect to their actual genre (row). The diagonal refers to the percentage of correct predictions.

561x709mm (39 x 39 DPI)

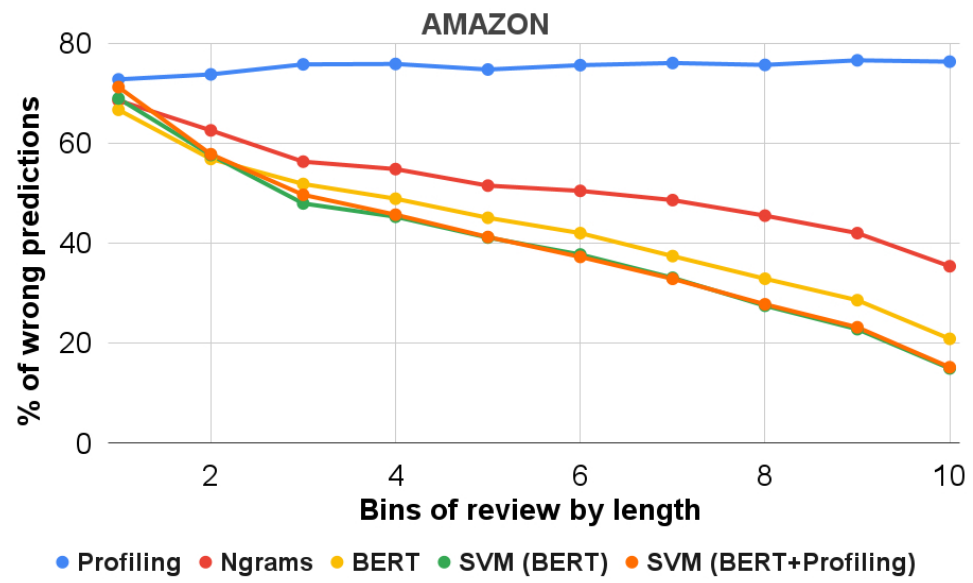


26 For the AGC task, the percentage of incorrect predictions of each model at different intervals of review
27 length.

28 332x198mm (72 x 72 DPI)

29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



For the AGC task, the percentage of incorrect predictions of each model at different intervals of review length.

332x198mm (72 x 72 DPI)