Unmasking the Wordsmith: Revealing Author Identity through Reader Reviews

Chiara Alzetta[1], Felice Dell'Orletta[1], Chiara Fazzone[1], Alessio Miaschi[1] and Giulia Venturi[1]

[1]ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy

Abstract

Traditional genre-based approaches for book recommendations face challenges due to the vague definition of genres. To overcome this, we propose a novel task called *Book Author Prediction*, where we predict the author of a book based on user-generated reviews' writing style. To this aim, we first introduce the 'Literary Voices Corpus' (LVC), a dataset of Italian book reviews, and use it to train and test machine learning models. Our study contributes valuable insights for developing user-centric systems that recommend leisure readings based on individual readers' interests and writing styles.

Keywords

Book Author Prediction, Italian reviews, stylistic analysis, user-generated book reviews

1. Introduction and Background

Reading for pleasure is currently experiencing a significant decline, as evidenced by surveys indicating that leisure reading has reached an unprecedented low[1]. Book recommender systems have been proposed as a valuable tool to promote the practice of reading for pleasure [1]. These systems provide personalized suggestions and aid users in navigating the vast array of available literary works [2]. Their integration into e-commerce services has long been explored, as it benefits both sellers and consumers [3].

Typically integrated with online platforms, book recommender systems rely on the history of users to predict their future interests and provide recommendations based on the literary genre or authors that users have previously engaged with. While recommending the other books by an author that the reader enjoyed is trivial, suggesting books belonging to the same genre remains a complex area of study, particularly concerning literary novels [4]. This is mostly due to the fact that the notion of *genre* represents a quite heterogeneous object of study due to multiple factors [5]. In fact, the same book can be assigned to more than one literary genre either on the same reading platform or across diverse platforms. Accordingly, various approaches have been proposed to automatically identify literary genres using book content [6, 7, 8], titles or summaries [9], and even cover designs [10]. Nevertheless, these models often face challenges when book content is inaccessible due to licensing restrictions.

Consequently, an alternative and promising line of research on book recommender systems involves leveraging user reviews as a valuable source of information for generating recommendations. Analyzing reviews allows for a unique perspective on books from the viewpoint of their readers, without requiring access to their content.

Reviews offer valuable insights into readers' opinions and preferences, and they have been effectively utilized to predict trends in the book market [11, 12, 13, 14, 15]. There are few attempts to exploit user reviews also for literary genre identification. These include [16] and [17] for English and Portuguese book reviews respectively. We have also contributed to this line of research by focusing on Italian book reviews [18]. In our previous work, we demonstrated how book reviews published by amateur readers on two social reading platforms, namely Amazon and Goodreads, can be exploited to automatically identify the genre of the reviewed book.

Building upon our prior investigations, our current research aims to explore whether the writing style of user-generated reviews, analyzed in terms of lexical and (morpho-)syntactic characteristics, can serve as a reliable source of information also to predict the author of a reviewed book. We started from the assumption that the vague definition of literary genres might make recommendations based on related authors more effective than genre-based approaches. To this end, inspired by the literature on Authorship Attribution [19], we introduced a novel task named *Book Author Prediction*. We tackle the problem as a supervised classification task, where the objective is to predict the author of a given book from a set of potential candidates. It is important to note that, unlike the traditional Authorship Attribution task, our information source consists of user-generated reviews

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

chiara.alzetta@ilc.cnr.it (C. Alzetta); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); chiara.fazzone@ilc.cnr.it (C. Fazzone); alessio.miaschi@ilc.cnr.it (A. Miaschi); giulia.venturi@ilc.cnr.it (G. Venturi)

[1]See https://www.istat.it/it/archivio/284591, https://literacytrust.org.uk/research-services/annual-literacy-survey/

rather than the books authored by the novelists themselves. This distinction adds a layer of complexity to the task, making it particularly challenging and novel in its approach. As a crucial step towards this objective, we introduce a novel dataset of Amazon[2] and Goodreads[3] book reviews, the 'Literary Voices Corpus' (LVC). The dataset successfully served in diverse experimental settings we explored in this work aimed at training and testing pretrained and traditional machine learning models, that use different configurations of lexical and (morpho-)syntactic features, to accomplish the new prediction task.

The work presented in this study falls within the context of collective efforts to foster the habit of reading and enlarge the readership across different target audiences[4]. Among these initiatives, LettERE (Letture pER TE) is a project that aims to encourage and promote the practice of reading by creating a reading recommendation system that provides personalised recommendations tailored to the reader's language skills and interests (see Acknowledgements). In this regard, the research presented in this paper contributes significantly to the LettERE project's objectives by showing that user-generated reviews can be effectively used to identify readers sharing common interests and ultimately provide personalised book recommendations.

The remainder of the paper is organised as follows. Section 2 presents LVC, the novel collection of Italian book reviews referring to the books of six popular authors. Section 3 introduces the Book Author Prediction task and details the methodology and models exploited in this work to address it. Section 4 presents the results of our experiments. Finally, Section 5 offers conclusions and outlines potential future research directions.

2. The Literary Voices Corpus

We performed our experiments on the 'Literary Voices Corpus' (LVC), which encompasses a collection of book reviews in Italian published on two leading platforms for Digital Social Reading (DSR), Amazon Books and Goodreads and covering the work of several authors of fiction novels.[5] This corpus is a spin-off of the 'A Good Review' corpus, which we introduced in [18]. The LVC corpus is aimed at being representative of two different approaches to writing book reviews, a diversity specific to the peculiarities of the two platforms. In fact, 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. Goodreads reviews are typically exploited to predict the orientation of the book market [11, 13], to map reading preferences across various communities of users [20], as well as to analyze the linguistic style adopted by readers to describe their reading experiences [21, 22]. Conversely, reviews posted on Amazon Books have mostly been investigated within marketing and buyers' behaviour studies, often relying on sentiment analysis [23, 24, 25].

When building LVC, we first chose popular novelists in order to acquire a diverse but rich collection of reviews from amateur readers. These are J.K. Rowling, Stephen King, J.R.R. Tolkien, Jane Austen, Sarah J. Maas, and Dan Brown.[6]

Since *literary genre* is not a monolithic notion [4], the books of these authors traverse multiple genres. For example, King's repertoire encompasses horror, thriller, and science-fiction, while Maas's fantasy novels also incorporate a substantial element of romance. Then, we extracted the reviews for their respective books from the 'A Good Review' corpus and we integrated the set with new books if necessary using the ISBN number of a book to unambiguously identify it on Amazon and Goodreads and to collect its reviews written in Italian. This was done to reach a minimum of 1,100 reviews per novelist from Goodreads and 800 reviews from Amazon. While we successfully obtained the desired number of reviews for most authors, we encountered challenges for Austen and Maas on Amazon. Nonetheless, the number of reviews collected for these authors can still be considered reasonably comparable to the desired amount. The statistics of the final LVC dataset are reported in Table 1.

As can be noted, the two portions of the dataset (i.e., Amazon and Goodreads) are quite different in terms of the length of a single review. This difference arises in part from the lower number of reviews collected from Amazon, but mostly from the comparatively greater length of Goodreads reviews in terms of sentences and tokens. Thus, achieving a balanced number of reviews across authors does not correspond to an equal number of tokens. Furthermore, we notice a tendency to produce longer reviews among the readers of certain authors, such as King, Maas, or Austen, on both platforms. This represents one of the first general characterization of the diversity across literary voices we collected.

3. Book Author Prediction

The novel task of Book Author Prediction consists of predicting the author of a book from the readers' reviews. We explored the performance on the task of a suite of machine learning algorithms that vary with re-

[2] https://www.amazon.it

[3] https://www.goodreads.com

[4] See for instance: https://www.regione.toscana.it/-/un-patto-per-la-lettura.

[5] The LVC corpus is freely available under request for research purposes.

[6] The complete list of books whose reviews in Italian have been included in LVC can be found in Appendix A.

	Rowling	King	Tolkien	Austen	Maas	Brown	All
				Goodreads			
Books	6	8	7	7	6	7	41
Reviews	1,100	1,100	1,100	1,100	1,100	1,100	6,600
Sentences Total	5,951	7,479	6,224	6,914	11,447	5,151	43,166
Tokens Total	155,653	202,027	180,680	214,921	302,687	129,684	1,185,652
Avg Sentences per Review	5.41	6.80	5.65	6.28	10.40	4.68	6.54
Avg Tokens per Review	141.50	183.66	164.25	195.38	275.17	117.89	179.64
				Amazon			
Books	6	8	6	7	5	7	39
Reviews	800	800	800	749	653	800	4,602
Sentences Total	1,712	3,525	2,695	2,326	3,961	2,422	16,641
Tokens Total	21,899	69,078	48,275	40,875	81,668	40,719	302,514
Avg Sentences per Review	2.14	4.40	3.36	3.10	6.06	3.03	3.61
Avg Tokens per Review	27.37	86.34	60.34	54.57	125.06	50.89	65.73

Table 1
Literary Voices Corpus statistics.

Raw text
Number of sentences and tokens
Average tokens per sentence and average characters per token
Vocabulary Richness
Type/Token Ratio for words and lemmas (first 100/200 tokens)
Morphosyntactic information
Distibution of UD POS
Lexical density
Inflectional morphology
Distribution of lexical verbs and auxiliaries for inflectional categories (tense, mood, person, number)
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Average verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Average depth of the whole syntactic trees
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Average clause length
Relative order of elements
Distribution of subjects and objects in post- and pre-verbal position
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Distribution of subordinates in post- and pre-principal clause position

Table 2
Linguistic features acquired from book reviews.

spect to the architecture and features used for training (see Section 3.1). The models leverage a wide spectrum of text properties acquired from the reviews of increasing informativeness, which range from n-grams of words to stylistic features (Section 3.2), up to contextual sentence representations of Neural Language Models. For all models, we adopted a 5-fold cross-validation approach for training and testing. The train and test sets always contain reviews of different books, thus increasing the complexity of the classification tasks. Note that, considering the high discriminative power of proper nouns in this classification scenario, we performed the linguistic analysis of reviews and sanitized the text [26] by masking all tokens marked as proper nouns (POS = PROPN).

3.1. Models

Linear Support Vector Machine We define two LinearSVM models, referred to as 'Profiling' and 'Ngrams' models. The former takes the set of linguistic characteristics described in Sec. 3.2. Ngrams exploits lexical information since it uses as input feature a simple contiguous sequence of *n* words acquired from the reviews (i.e. n-grams, with *n* equal to 1, 2, and 3).

Neural Language Model We relied on the Italian pretrained version of the BERT model (12 layers, 768 hidden units) [27][7], which was pretrained using the Italian Wikipedia and the Italian portion of the OPUS corpus [28], a multilingual collection of translated open source documents available on the Internet, and fine-tuned on the Book Author Classification task.

LinearSVM + NLM We combined the previous models into a classifier based on LinearSVM and trained using the internal representations of the BERT model fine-tuned on the author classification tasks. We refer to this model as *SVM (BERT)*. *SVM (BERT+Profiling)* is an additional LinearSVM model trained using both the fine-tuned representations produced by BERT and Profiling-UD features. The BERT representations used as input features of the SVM model were computed by averaging the embeddings of all the tokens in each review.

Baselines We compared the performance of the above models against a random uniform classifier, i.e. a model that uniformly generates random predictions for each author.

[7] https://huggingface.co/dbmdz/bert-base-italian-cased

	Rowling	King	Tolkien	Austen	Maas	Brown	All
Model				**Goodreads**			
Baseline	0.19	0.15	0.16	0.18	0.15	0.16	0.16
Profiling	0.21	0.18	0.26	0.27	0.40	0.25	0.26
Ngrams	0.42	0.36	0.46	0.51	0.46	0.44	0.44
BERT	**0.69**	**0.70**	**0.72**	**0.79**	**0.73**	**0.74**	**0.73**
SVM (BERT)	0.44	0.51	0.55	0.58	0.57	0.56	0.54
SVM (BERT + Profiling)	0.46	0.50	0.51	0.54	0.56	0.57	0.52
Average	0.44	0.45	0.50	**0.54**	**0.54**	0.51	0.50
				Amazon			
Baseline	0.16	0.15	0.17	0.16	0.16	0.14	0.16
Profiling	0.38	0.18	0.27	0.17	0.32	0.22	0.26
Ngrams	0.44	0.35	0.40	0.38	0.58	0.39	0.42
BERT	**0.57**	**0.60**	**0.56**	**0.64**	**0.72**	**0.61**	**0.61**
SVM (BERT)	0.39	0.40	0.45	0.45	0.63	0.43	0.46
SVM (BERT+Profiling)	0.41	0.42	0.39	0.46	0.56	0.36	0.43
Average	0.44	0.39	0.41	0.42	**0.56**	0.40	0.44

Table 3
Results of book author prediction on Goodreads and Amazon reviews.

3.2. Linguistic Features

To model the linguistic properties of the reviews, we relied on a set of 150 linguistic features. These features correspond to specific aspects of the document structure and were derived using Profiling-UD [29], a web-based tool conceived to linguistically profile multilingual texts by relying on the Universal Dependencies (UD) formalism [30]. The features encompass 9 dimensions of document structure, which are detailed in Table 2. They range from morpho-syntactic and inflectional properties to more complex aspects of sentence structure, such as the depth of the syntactic tree. Other features pertain to the structure of sub-trees and include the order of subjects and objects in relation to the verb, as well as the use of subordination.

4. Results

Table 3 presents the classification accuracies for the task of Book Author Prediction. Notably, all models outperformed the random uniform baseline on both Amazon and Goodreads. Upon closer examination of the models, we notice that lexical information has more discriminative power than linguistic properties in the task. As proof, consider the global and author-level scores obtained by the *Profiling* model compared to the *Ngram* and, most notably, the *BERT* models. Interestingly, using the fine-tuned BERT representations as input features for the SVM classifier (*SVM (BERT)*) yielded lower results than simply using pre-trained BERT, and the results are comparable – or lower – when combining contextualized representations with linguistic features (*SVM (BERT+Profiling)*).

Comparing the two platforms, Goodreads reviews exhibit on average higher accuracy scores overall. This is possibly due to a typical trait of commercial platforms like Amazon, whose reviews frequently encompass aspects beyond the book's content, such as parcel delivery or the edition's book cover. These topics cause the reviews to be quite standardised, thus more difficult to discriminate. Conversely, Goodreads reviews primarily focus on the book's content possibly containing a larger amount of stylistic elements which help the automatic classification. This trend holds also when classifying individual authors, except Rowling for the *Profiling* and *Ngrams* models.

When looking at the results obtained for individual authors, Sara J. Maas turned out to be the most accurately predicted author on both platforms, considering the average scores across all models. However, upon closer inspection of the results obtained with the top-performing model (*BERT*), we observe that while Maas remains the most accurately identified author in Amazon reviews, the reviews of Jane Austen's books exhibit the highest level of distinctiveness on Goodreads.

4.1. Discussion

To take a closer look at the classification results, Fig. 1 reports the confusion matrices with the percentage of the predictions made by all models in the Book Author Prediction task. This complements the classification results by showing which authors are more confusing and which are the most wrongly classified ones.

In general, we observe that as the model performance improves, the matrices become less sparse, regardless of the platform. This means that when the correct author is predicted most of the time, the erroneous predictions are distributed quite evenly among all possible authors.

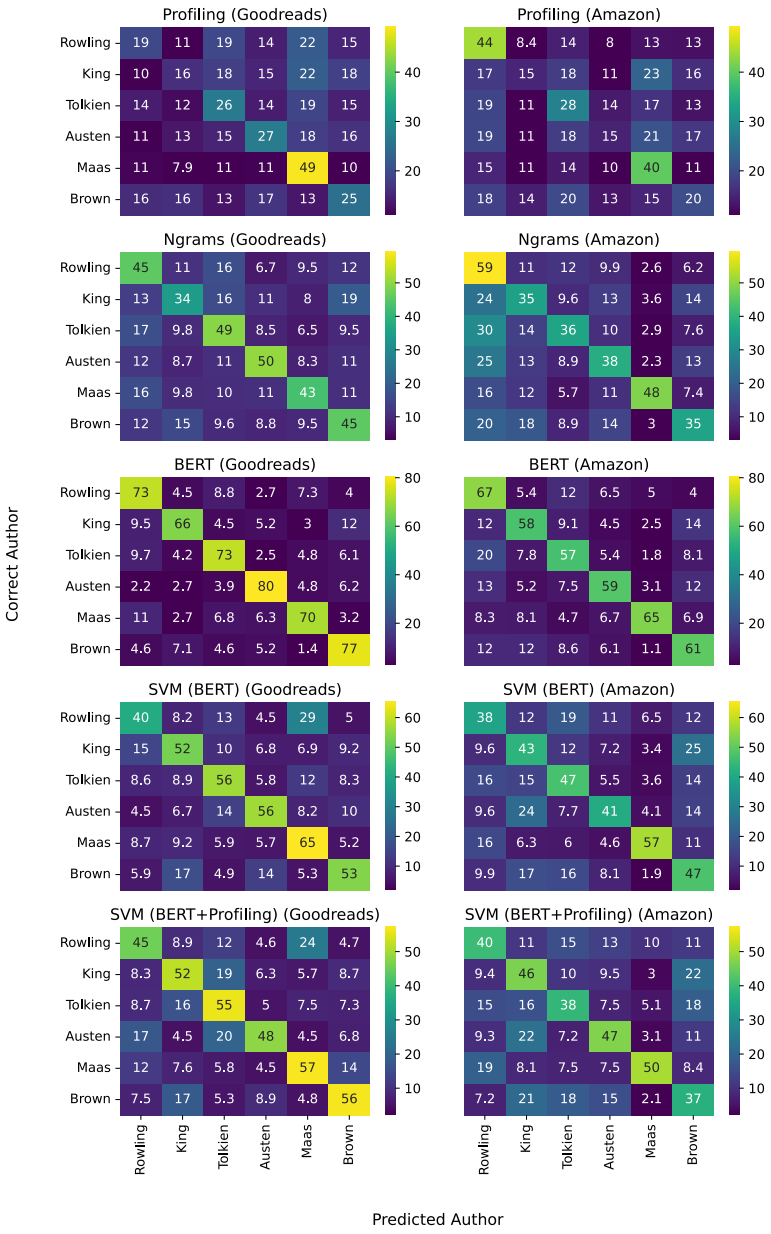

Figure 1: Confusion matrices of the classification task for all models: cells report the percentage of reviews automatically assigned to an author by each classification model (column) with respect to their actual author (row).

Consider, for instance, the matrices obtained from the analysis of *BERT* and compare them with the matrices referring to the *Profiling* and *Ngrams* models, which yield the most sparse matrices.

Notable differences arise in the distribution of predicted authors across the two platforms. For instance, when considering the *Profiling* model applied to Goodreads reviews, we observe that Maas is the most frequently predicted author, leading to other authors' books being frequently misclassified as Maas's works. Notably, the reviews of *It* by King and of the fourth book from the Harry Potter saga by Rowling are often incorrectly assigned to Maas. The content of these books, at the crossroads between the fantasy and horror genres,

may contribute to the model confusion. However, the most influencing factor to the *Profiling* model predictions appears to be the review length. On Goodreads, reviews of King's and Rowling's books that are longer than 150 tokens are wrongly classified as referring to Maas in over 40% of cases. On Amazon, we observe an opposite tendency, but for a different author: when a review has less than 10 tokens, the model assigns the review to Rowling in around 60% of cases.

The analysis of the feature rankings[8] produced by the classifiers trained on both Amazon and Goodreads reviews confirms the importance of review length for the *Profiling* model. Indeed, features that capture structural properties are particularly relevant for the model: the use of subordination (*subordinate_dist*) is crucial for classifying Rowling's and King's reviews on Goodreads, as they exhibit respectively the lowest and highest use of subordinate clauses. Conversely, on Amazon, the average number of verb dependents (*verb_edges*) and the distribution of function words (namely, conjunctions, auxiliary verbs and determiners) are discriminative for Rowling, Tolkien, and Maas.

For what concerns the *Ngram* model, the feature ranking consists of the n-grams employed by the model ordered by relevance for book author classification purposes on Amazon and on Goodreads. Quite expectedly, the analysis of the top 100 most relevant n-grams reveals that, on Amazon, parcel delivery is a highly referenced topic (e.g. '*tempi previsti*', expected timing, and '*ben confezionato*', well packaged), especially among the readers of Tolkien and Rowling, which have the most similar n-gram rankings (Spearman correlation score = 0.235, $p < 0.05$). The two authors are the most frequently confused by the model, especially for what concerns the reviews of Tolkien's 'The Hobbit' and 'The Silmarillion', wrongly classified as referring to Rowling's books. Indeed, it is possible that the two authors attract a similar readership interested in books involving intricate mythologies, and that feature multi-dimensional characters with strengths, flaws, and internal struggles. Such closeness between the Amazon reviews of these authors is captured also by the BERT model which, although performing better than other models on the task, seems quite confused by the reviews of the same Tolkien books.

On Goodreads reviews, where parcel delivery is not relevant, the most impactful n-grams tend to revolve around book appreciation (e.g., *'ho apprezzato'*, I appreciated; *'lettura piacevole'*, pleasant reading; *'non mi aspettavo'*, I did not expect) or plot (*'il maghetto'*, the little wizard; *'signore di'*, lord of; *'chiesa'*, church; *'di epoca'*, historical; *'drago'*, dragon; *'di vampiri'*, of vampires). Therefore, it is not surprising to see that King's reviews are most frequently misclassified as referring to Brown's work, also by the BERT model. Both authors, despite their differences, are known for building suspense and tension in their narratives and incorporating detailed historical settings and psychological aspects into their work.

The classification of Goodreads review performed by the *SVM (BERT)* and *SVM (BERT + Profiling)* models highlight author commonalities that did not emerge so strongly with other models. The reviews of Rowling's books, for instance, are frequently wrongly classified as referring to Maas's work. Both authors are known for their contributions to popular literature, particularly in the genres of fantasy and young adult fiction, which attract a readership interested in exploring themes of personal growth and self-discovery through the characters' coming-of-age journeys.

Overall, no particular author appears to be systematically confused by all models. This finding is particularly interesting from our perspective since it shows that using user-generated reviews as an information source allows to successfully address the Book Author Prediction task. It suggests that books authored by different novelists attract readers who are interested in similar topics and also adopt similar communication strategies in their writing. It also implies that the proposed methodology could have a positive impact on the development of user-centric book recommender systems.

5. Conclusions

This paper has explored an innovative approach that leverages user reviews as a source of information for Book Author Prediction. Building upon our prior work, we introduced a novel dataset of Amazon and Goodreads book reviews, LVC, which has been used for training and evaluating machine learning models addressing the novel book author prediction task.

Our findings highlight the challenging nature of predicting the author of a novel from a reader's review. However, the analysis of erroneous predictions pointed us to cases of books sharing a similar readership. This observation supports the intuition that user-generated reviews can effectively serve as a basis for personalized book recommendations. By analyzing reviews, we gained insights into readers' preferences beyond the writing style of the book's author, opening up new avenues for more tailored and user-centric recommendations.

Moving forward, this research could be expanded by investigating the impact of exploiting user judgments as an additional feature for classification. Furthermore, the sentiment expressed by readers about a book, whether positive or negative, could be leveraged to validate and fine-tune personalized recommendations.

[8] See Appendix B and C.

Acknowledgments

We thank the "Letture pER TE" (LettERE) project (2022-2024) funded by Regione Toscana (Progetti Congiunti di Alta Formazione – POR FSE 2014-2020 Investimenti a favore della crescita e dell'occupazione) in collaboration with M.E.T.A. Srl company.

References

[1] H. Alharthi, D. Inkpen, S. Szpakowicz, Authorship identification for literary book recommendations, in: Proceedings of the 27th International Conference on Computational Linguistics (COLING), ACL, 2018, pp. 390–400.

[2] H. Alharthi, D. Inkpen, S. Szpakowicz, A survey of book recommender systems, Journal of Intelligent Information Systems 51 (2018) 139–160.

[3] J. B. Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, in: Proceedings of the 1st ACM conference on Electronic commerce, 1999, pp. 158–166.

[4] J.-M. Schaeffer, Qu'est-ce qu'un genre littéraire?, Seuil, 1989.

[5] D. Biber, S. Conrad, Genre, Register, Style, Cambridge University Press, 2009.

[6] L. Shamir, UDAT: Compound quantitative analysis of text using machine learning, Digital Scholarship in the Humanities 36 (2020) 187–208.

[7] Rahul, Ayush, D. Agarwal, D. Vijay, Genre classification using character networks, in: Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 216–222.

[8] J. Worsham, J. Kalita, Genre identification and the compositional effect of genre in literature, in: Proceedings of the 27th International Conference on Computational Linguistics (COLING), ACL, 2018, pp. 1963–1973.

[9] E. Ozsarfati, E. Sahin, C. J. Saul, A. Yilmaz, Book genre classification based on titles with comparative machine learning algorithms, in: Proceedings of 2019 4th International Conference on Computer and Communication Systems (ICCCS), IEEE, 2019, pp. 14–20.

[10] P. Buczkowski, A. Sobkowicz, M. Kozlowski, Deep learning approaches towards book covers classification, in: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM), SCITEPRESS-Science and Technology Publications, 2018, pp. 309–316.

[11] K. Wang, X. Liu, Y. Han, Exploring Goodreads reviews for book impact assessment, Journal of Informetrics 13 (2019) 874–886.

[12] G. Aerts, T. Smits, P. W. Verlegh, How online consumer reviews are influenced by the language and valence of prior reviews: A construal level perspective, Computers in Human Behavior 75 (2017) 855–864.

[13] S. K. Maity, A. Panigrahi, A. Mukherjee, Analyzing social book reading behavior on Goodreads and how it predicts Amazon best sellers, Influence and Behavior Analysis in Social Networks and Social Media (2019) 211–235.

[14] S. Dimitrov, F. Zamal, A. Piper, D. Ruths, Goodreads versus Amazon: the effect of decoupling book reviewing and book selling, in: Proceedings of International AAAI Conference on Web and Social Media (ICWSM), volume 9, 2015, pp. 602–605.

[15] M. Thelwall, Reader and author gender and genre in Goodreads, Journal of Librarianship and Information Science 51 (2019) 403–430.

[16] M. Saraswat, Leveraging genre classification with rnn for book recommendation, International Journal of Information Technology (2022) 1–6.

[17] C. Scofield, M. O. Silva, L. de Melo-Gomes, M. M. Moro, Book genre classification based on reviews of portuguese-language literature, in: Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR), 2022, pp. 188–197.

[18] C. Alzetta, F. Dell'Orletta, A. Miaschi, E. Prat, G. 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 Forthcoming (2023).

[19] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556.

[20] K. Bourrier, M. Thelwall, The social lives of books: Reading victorian literature on Goodreads, Journal of Cultural Analytics 5 (2020) 12049.

[21] B. Driscoll, D. Rehberg Sedo, Faraway, so close: Seeing the intimacy in Goodreads reviews, Qualitative Inquiry 25 (2019) 248–259.

[22] L. Nuttall, C. Harrison, Wolfing down the twilight series: Metaphors for reading in online reviews, Contemporary media stylistics (2020) 35–60.

[23] K. Kaur, T. Singh, Impact of online consumer reviews on Amazon books sales: Empirical evidence from india, Journal of Theoretical and Applied Electronic Commerce Research 16 (2021) 2793–2807.

[24] F. Chiavetta, G. L. Bosco, G. Pilato, A lexicon-based approach for sentiment classification of Amazon books reviews in Italian language, in: International Conference on Web Information Systems and Technologies (WEBIST), volume 3, Scitepress, 2016, pp. 159–170.

[25] K. Srujan, S. Nikhil, H. Raghav Rao, K. Karthik, B. Harish, H. Keerthi Kumar, Classification of Amazon book reviews based on sentiment analysis, in: Information Systems Design and Intelligent Applications, Springer, 2018, pp. 401–411.
[26] V. Vasudevan, A. John, A review on text sanitization, International Journal of Computer Applications 95 (2014).
[27] T. Wolf, L. Debut, V. Sanh, alii, Transformers: State-of-the-art natural language processing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2020, pp. 38–45.
[28] J. Tiedemann, L. Nygaard, The OPUS corpus - parallel and free, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), ELRA, 2004.
[29] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), ELRA, 2020, pp. 7147–7153.
[30] M. C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational linguistics 47 (2021) 255–308.

A. Books of the Literary Voices Corpus

Author	Book
Jane Austen	Emma
	Lady Susan
	Mansfield Park
	Northanger Abbey
	Persuasion
	Persuasion
	Pride and Prejudice
	Sense and Sensibility
Dan Brown	Angels and Demons
	Deception Point
	Digital Fortress
	Inferno
	Origin
	The Da Vinci Code
	The Lost Symbol
Sarah J. Maas	A Court of Mist and Fury
	A Court of Frost and Starlight
	A Court of Wing and Ruin
	A Court of Silver Flames
	Throne of Glass
J.K. Rowling	Harry Potter and the Chamber of Secrets
	Harry Potter and the Goblet of Fire
	Harry Potter and the Half Blood Prince
	Harry Potter and the Order of the Phoenix
	Harry Potter and the Prisoner of Azkaban
	Harry Potter and the Sorcerer's Stone
J.R.R. Tolkien	The Fellowship of the Ring
	The Children of Húrin
	The Hobbit
	The Return of the King
	The Silmarillion
	The Two Towers
Stephen King	Salem's Lot
	Carrie
	Doctor Sleep
	It
	Misery
	Mr. Mercedes
	Pet Sematary
	The Shining

Table 4
List of the books present in the LVC.

B. Feature ranking Profiling Model (Goodreads)

	Dan Brown		J.K. Rowling		J.R.R. Tolkien	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	ttr_form_chunks_100	0.22	subordinate_post	67.91	ttr_lemma_chunks_100	0.26
2	ttr_lemma_chunks_100	0.20	subordinate_dist_1	58.66	ttr_form_chunks_100	0.30
3	upos_dist_AUX	4.46	subordinate_pre	11.46	aux_tense_dist_Pres	71.32
4	avg_prepositional_chain_len	0.93	dep_dist_orphan	0.00	ttr_form_chunks_200	0.15
5	dep_dist_aux	2.54	verbs_form_dist_Part	26.69	ttr_lemma_chunks_200	0.13
6	upos_dist_DET	12.87	ttr_form_chunks_100	0.24	n_prepositional_chains	7.29
7	dep_dist_cop	1.57	avg_prepositional_chain_len	0.86	n_tokens	164.25
8	prep_dist_2	10.46	upos_dist_ADP	11.04	upos_dist_AUX	4.84
9	prep_dist_1	68.65	subordinate_dist_2	15.22	upos_dist_ADP	12.02
10	dep_dist_det	12.06	dep_dist_mark	2.69	dep_dist_orphan	0.00
11	ttr_form_chunks_200	0.10	upos_dist_SCONJ	1.54	upos_dist_DET	14.35
12	ttr_lemma_chunks_200	0.09	ttr_lemma_chunks_100	0.21	aux_mood_dist_Ind	73.27
13	dep_dist_flat:name	1.38	verb_edges_dist_1	15.34	dep_dist_aux	2.60
14	avg_verb_edges	2.39	aux_tense_dist_Pres	67.71	aux_tense_dist_Imp	5.79
15	prep_dist_3	1.02	avg_subordinate_chain_len	1.08	dep_dist_case	10.55
16	dep_dist_cc	3.42	verb_edges_dist_2	25.27	dep_dist_cop	1.90
17	dep_dist_flat:foreign	0.08	dep_dist_case	9.77	dep_dist_mark	2.64
18	upos_dist_NUM	0.77	verb_edges_dist_3	23.72	verbs_form_dist_Part	28.82
19	upos_dist_PROPN	4.68	verbs_form_dist_Fin	38.53	dep_dist_flat:name	0.62
20	upos_dist_CCONJ	3.43	upos_dist_AUX	4.97	aux_num_pers_dist_Sing+3	52.14

	Jane Austen		Sarah J. Maas		Stephen King	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	ttr_lemma_chunks_200	0.15	ttr_form_chunks_200	0.25	upos_dist_CCONJ	3.29
2	ttr_form_chunks_200	0.18	ttr_lemma_chunks_200	0.22	dep_dist_cc	3.29
3	upos_dist_CCONJ	3.88	verbs_form_dist_Fin	38.74	avg_prepositional_chain_len	0.98
4	verbal_head_per_sent	3.35	verbs_form_dist_Part	30.77	ttr_form_chunks_200	0.17
5	avg_prepositional_chain_len	0.99	verb_edges_dist_2	27.51	ttr_lemma_chunks_200	0.15
6	n_tokens	195.38	verb_edges_dist_1	12.55	prep_dist_2	10.28
7	dep_dist_cc	3.87	verb_edges_dist_3	26.86	prep_dist_1	74.51
8	verbs_form_dist_Fin	38.72	verbs_form_dist_Inf	21.37	subordinate_post	76.19
9	tokens_per_sent	29.92	aux_tense_dist_Past	5.30	dep_dist_orphan	0.00
10	ttr_lemma_chunks_100	0.30	avg_prepositional_chain_len	0.95	prep_dist_3	0.89
11	prep_dist_1	72.48	n_prepositional_chains	9.46	subordinate_dist_1	66.18
12	ttr_form_chunks_100	0.34	verb_edges_dist_4	16.36	tokens_per_sent	26.85
13	n_sentences	6.29	aux_tense_dist_Pres	75.14	n_tokens	183.66
14	dep_dist_advmod	7.51	prep_dist_1	74.15	aux_tense_dist_Pres	72.50
15	prep_dist_2	10.92	verbal_head_per_sent	3.60	ttr_lemma_chunks_100	0.31
16	verb_edges_dist_2	27.55	aux_form_dist_Part	5.03	avg_verb_edges	2.54
17	verb_edges_dist_3	26.57	prep_dist_2	9.36	subordinate_pre	12.63
18	upos_dist_ADV	8.00	n_tokens	275.17	verbal_head_per_sent	3.22
19	upos_dist_AUX	4.79	aux_mood_dist_Ind	78.48	dep_dist_case	10.52
20	dep_dist_case	10.32	verb_edges_dist_5	6.62	upos_dist_ADP	12.03

Table 5
Top 20 ranked features by the Profiling model for the classification of each author on Goodreads. Average values of the linguistic features are also reported (columns *Avg*).

C. Feature ranking Profiling Model (Amazon)

	Dan Brown		J.K. Rowling		J.R.R. Tolkien	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	avg_subordinate_chain_len	0.98	ttr_form_chunks_200	0.01	upos_dist_AUX	4.57
2	dep_dist_cc	3.47	ttr_lemma_chunks_200	0.01	dep_dist_det	12.67
3	upos_dist_AUX	4.06	upos_dist_CCONJ	2.67	dep_dist_aux	2.23
4	upos_dist_CCONJ	3.46	dep_dist_cc	2.67	upos_dist_ADV	6.84
5	aux_tense_dist_Pres	62.35	upos_dist_AUX	3.99	ttr_lemma_chunks_100	0.07
6	dep_dist_aux	2.18	ttr_form_chunks_100	0.03	ttr_form_chunks_100	0.08
7	dep_dist_cop	1.63	dep_dist_aux	2.11	dep_dist_cop	1.95
8	subordinate_dist_2	12.89	verb_edges_dist_3	16.65	upos_dist_DET	13.35
9	lexical_density	0.57	verb_edges_dist_2	25.40	dep_dist_root	10.92
10	verbs_form_dist_Part	33.30	ttr_lemma_chunks_100	0.03	dep_dist_advmod	6.41
11	ttr_lemma_chunks_200	0.01	n_tokens	27.37	verb_edges_dist_2	29.13
12	upos_dist_DET	12.06	dep_dist_cop	1.57	verb_edges_dist_3	21.28
13	subordinate_dist_3	2.54	verb_edges_dist_4	6.61	ttr_lemma_chunks_200	0.02
14	aux_form_dist_Fin	63.30	verbs_form_dist_Inf	11.28	aux_tense_dist_Pres	64.68
15	subordinate_dist_1	61.25	lexical_density	0.65	verb_edges_dist_4	11.55
16	verbs_form_dist_Fin	34.92	aux_form_dist_Part	2.23	avg_verb_edges	2.14
17	upos_dist_PUNCT	10.96	verb_edges_dist_1	18.43	verbs_form_dist_Part	35.10
18	ttr_form_chunks_200	0.01	avg_verb_edges	1.59	dep_dist_case	10.77
19	verb_edges_dist_3	23.54	verbs_form_dist_Fin	21.86	ttr_form_chunks_200	0.02
20	dep_dist_case	10.13	aux_tense_dist_Past	2.27	verb_edges_dist_1	18.95

	Jane Austen		Sarah J. Maas		Stephen King	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	aux_tense_dist_Pres	59.15	verbs_form_dist_Part	33.77	dep_dist_det	11.73
2	upos_dist_AUX	4.30	ttr_lemma_chunks_200	0.09	dep_dist_cc	3.12
3	avg_verb_edges	2.13	ttr_form_chunks_200	0.10	upos_dist_CCONJ	3.19
4	upos_dist_DET	12.40	verbs_form_dist_Fin	33.45	upos_dist_DET	12.63
5	dep_dist_case	9.13	verb_edges_dist_2	30.79	ttr_form_chunks_200	0.05
6	upos_dist_ADP	10.77	lexical_density	0.54	ttr_lemma_chunks_200	0.05
7	dep_dist_cc	3.73	verb_edges_dist_1	15.47	avg_verb_edges	2.27
8	dep_dist_aux	2.36	verbs_form_dist_Inf	20.95	verbs_form_dist_Part	32.46
9	verbs_form_dist_Part	28.94	verb_edges_dist_3	23.55	aux_tense_dist_Pres	63.26
10	avg_subordinate_chain_len	0.90	dep_dist_flat	0.00	verbs_form_dist_Fin	34.79
11	dep_dist_det	11.54	upos_dist_DET	13.45	verbs_form_dist_Inf	19.15
12	upos_dist_CCONJ	3.73	upos_dist_NUM	0.55	lexical_density	0.55
13	dep_dist_cop	1.67	verb_edges_dist_4	12.00	ttr_lemma_chunks_100	0.12
14	verbs_form_dist_Fin	34.52	dep_dist_nummod	0.53	verb_edges_dist_1	15.93
15	aux_mood_dist_Ind	59.79	upos_dist_ADP	10.91	dep_dist_root	11.04
16	ttr_form_chunks_200	0.02	dep_dist_flat:name	0.31	upos_dist_AUX	4.45
17	ttr_lemma_chunks_200	0.02	verb_edges_dist_0	2.18	dep_dist_aux	2.44
18	aux_form_dist_Fin	61.29	avg_subordinate_chain_len	1.02	principal_proposition_dist	39.50
19	subordinate_dist_2	11.26	dep_dist_det	12.28	dep_dist_det:poss	0.68
20	aux_tense_dist_Imp	3.92	verbs_form_dist_Ger	2.63	dep_dist_flat:foreign	0.03

Table 6
Top 20 ranked features by the Profiling model for the classification of each author on Amazon. Average values of the linguistic features are also reported (columns *Avg*).