Mining Goodreads. A Digital Humanities Project for the Study of Reading Absorption

Simone Rebora[1, 2*], Moniek Kuijpers[1], Piroska Lendvai[1]

1 University of Basel, Switzerland

2 University of Verona, Italy

*Corresponding author: Simone Rebora simone.rebora@unibas.ch

Abstract

We present our method and interim results of the "Mining Goodreads" project, aimed at developing a computational approach to measure reading absorption in user-generated book reviews in English. A team of eight people (three supervisors and five annotators) have joined skills from the fields of empirical literary studies, natural language processing, and digital humanities, with the goal of producing a gold-standard annotated dataset and strengthening the theoretical framework of reading absorption. Annotation of more than 800 texts showed the difficulties in finding an agreement in the tagging of sentences. However, through more than one year of work in strict collaboration, the team reached some substantial improvements: inter-annotator agreement increased through seven annotation rounds, while machine learning approaches were applied on the annotated corpus, producing promising results.

keywords

reading absorption; social reading; empirical literary studies; machine learning; inter-annotator agreement

I THE PROJECT'S IDEA

The "Mining Goodreads" project is conceived as a computational expansion of empirical literary studies. Empirical studies frequently use methods such as interviews or questionnaires to test theories and verify hypotheses, but they can also involve technologies such as eye-tracking and fMRI scans (for a general introduction, see [Peer, Hakemulder, and Zyngier 2012]). In all cases, direct involvement of readers in experiments is required to investigate reading experiences and their effects.

One of the most researched topics in empirical literary studies is narrative absorption, understood as the sensation of being absorbed into a story (see [Hakemulder, Kuijpers, and Tan 2017]). [Kuijpers et al. 2014] developed the Story World Absorption Scale (SWAS), a questionnaire aimed at measuring different dimensions of absorption into fictional worlds of literature. The questionnaire is built upon a theorization that distinguishes four main dimensions: attention (focused attention on the text, reducing awareness of the self and the passing of time), transportation (the feeling of having traveled to the story world), emotional engagement, and mental imagery. A total of 18 statements express different facets of what it is like to feel absorbed in a story (e.g., "When I finished the story I was surprised to see that time had gone by so fast", or "I could imagine what it must be like to be in the shoes of the main character"). During experiments using the SWAS, participants are asked to read narratives and rate their agreement with each of the 18 statements on a 7-point Likert scale. Statistically significant trends among readers' answers can then be used to quantify

intrinsically absorbing properties of texts. The SWAS has been empirically validated and used in multiple studies (e.g. [Bálint et al. 2016; Hartung et al. 2017; Kuzmičová et al. 2017]).
[Rebora, Lendvai, and Kuijpers 2018] showed a possible alternative use of the SWAS, building on the fact that multiple sentences in reviews published on the *Goodreads* platform [http1] overlap semantically and conceptually with SWAS statements. For example, a reviewer writes: "I'm so absorbed in the world Martin produced out of his wits" (a sentence that matches with the SWAS statement "I felt absorbed in the story"); another reviewer expresses her identification with the main character: "I went through all the emotional ups and downs right along with her" (matching with "I felt how the main character was feeling"). This phenomenon offers the possibility of using the SWAS without directly involving readers in experiments: an estimate of the absorbing properties of a book (or of a literary genre) can be inferred directly from its reviews. Possible noisiness and unreliability of reviews is countered by the fact that *Goodreads* hosts about 90 million reviews [http2]: a big data repository that can be studied from a "distant reading" perspective. Inevitably, such a wide repository requires the application of computational methods for its analysis.

II THE PROJECT'S STRUCTURE

2.1 People

The "Mining Goodreads" project, funded by the Swiss National Science Foundation in the "Digital Lives" funding scheme (grant number 10DL15_183194), involves a team of eight people. The three supervisors embody the three main disciplines involved: empirical studies, which provides the theoretical framework on reading absorption; natural language processing, which develops methods to automatically identify and retrieve absorption statements; and digital humanities, which mediates between the two by grounding the research in a "distant reading" perspective. At the core of the project is the work of five annotators, whose goal is that of generating annotations that will be adjudicated and consolidated into a ground truth dataset to train algorithms of different types. Once able to recognize absorption statements with an acceptable level of accuracy, these algorithms will scale up the analysis to millions of reviews.

2.2 Resources

A corpus of about 6 million reviews (amounting to more than 900 million tokens) has been generated by scraping the *Goodreads* website between 2018 and 2019. Titles were selected by focusing on 9 genres (as categorized by the *Goodreads* tagging system, which allows multiple genre assignments): general statistics are provided by Figure 1.
Due to property and privacy issues, we have provisionally decided against publicly sharing the corpus. However, new European directives are suggesting the introduction of significant exceptions in text and data mining for research purposes, e.g. the *Directive on Copyright in the Digital Single Market*. Some of these exceptions have been already included in national laws such as the *Urheberrechts-Wissensgesellschafts-Gesetz* in Germany, the country where our project started and where the entire scraping activity took place. We are currently evaluating the possibility of making the annotated corpus accessible under a specific license, after having complied with all legal and ethical requirements.

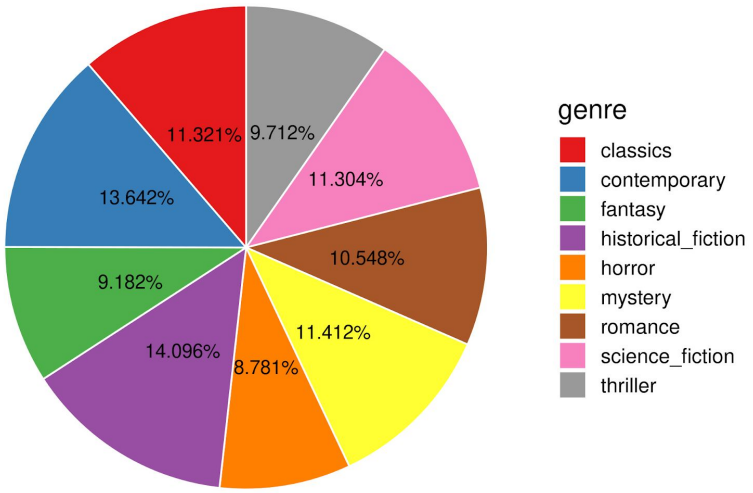

Figure 1. Proportions of genres in the corpus.

2.3 Procedure

Between March 2019 and May 2020, the five annotators have tagged a total of 830 reviews. After two months of training (getting acquainted with the absorption framework and with the annotation infrastructure), work has been split into seven annotation rounds: for each round, annotators were assigned a new batch of reviews to be annotated in parallel; at the end of each round, they met with the supervisors to discuss discrepancies in their annotation strategies. This procedure was aimed at improving inter-annotator agreement without directly interfering with the annotation work. Inter-round meetings proved fundamental also to strengthen (and eventually redefine) the theoretical framework of reading absorption. As shown by Table 1, amounts of annotated reviews gradually diminished at each round (offering the possibility to meet more frequently), while the number of tags increased substantially (mirroring a more precise distinction of the phenomena to be tagged).

Annotation round	Annotated reviews	Number of tags
1	180	6
2	200	12
3	150	80
4	60	145
5	90	145
6	75	145
7	75	145

Table 1. Number of tags and annotated reviews.

Tagsets were expanded by using a hierarchical structure, where all new tags can always be collapsed into a few, higher-level tags: **SWAS_specific**, for sentences that show direct similarity with the SWAS statements; **SWAS_related**, for sentences not included in the SWAS, but listed in a wider taxonomy of reading absorption [Bálint et al., 2016]; and **mention_SWAS**, for mentions of the SWAS concepts without reference to the actual reading experience of the user who wrote the review (i.e., "usually when I read a book, I like to be able to fully imagine what the world of the story looks like"). To these labels was also added a **Present/Absent** flag, for distinguishing sentences that explicitly confirm or negate absorption concepts.

Annotations were initially performed using the *brat* platform [Stenetorp et al. 2012], while from round 4 the *INCEpTION* platform [Klie et al. 2018] was adopted, which offered more advanced functionalities.

III PRELIMINARY RESULTS

3.1 Inter-annotator agreement

Figure 2 shows the evolution of Krippendorff's Alpha for the main tags in the seven rounds. As evident, there is a slight but steady improvement throughout the annotation process, that can be verified via the evolution of the "mean" and "all" scores: "mean" indicates the mean of the alpha scores for all of the tags (as it was not possible to calculate a single alpha score, because different tags could be assigned to the same sentences); "all" indicates the alpha score for a unique tag, obtained by checking if the sentence was annotated or not, independently from the assigned tag. In both cases, values move from fair (~0.2/0.4) to substantial agreement (~0.6/0.7). Among the high-level tags, SWAS_related_PRESENT reaches the highest values, while mention_SWAS_PRESENT scores the lowest, confirming the difficulty in recognizing absorption when no experiences of the I are mentioned.

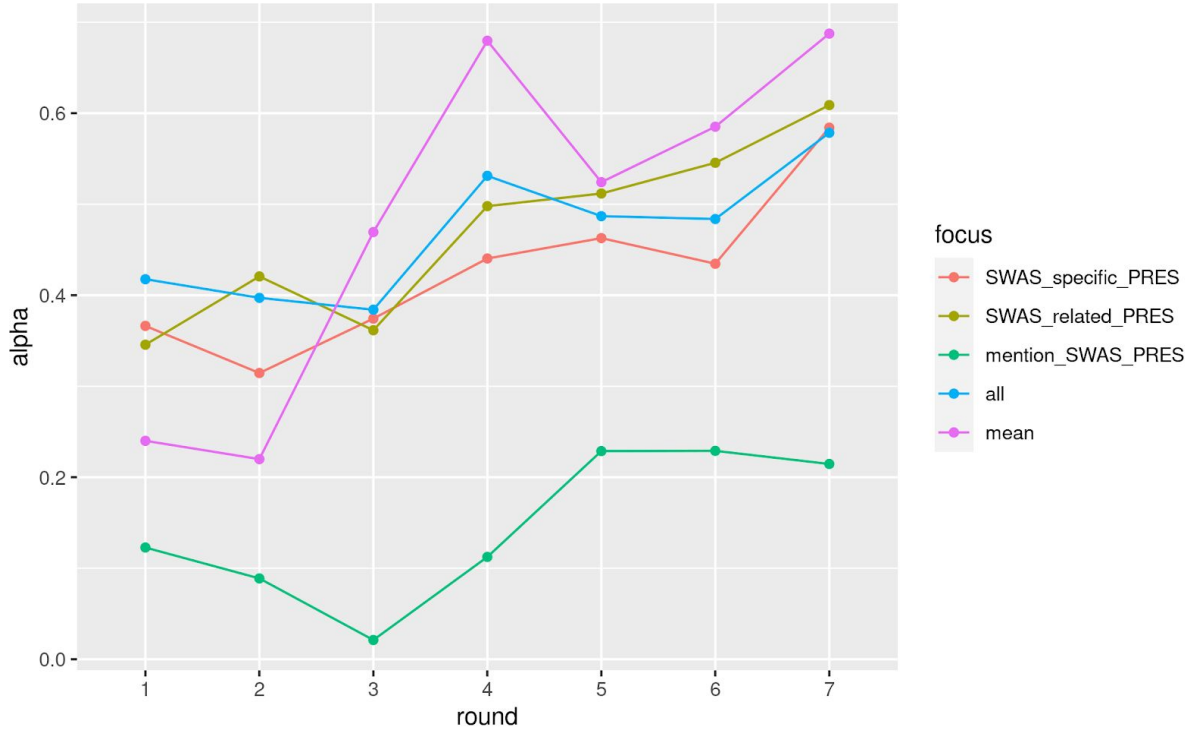

Figure 2. Inter-annotator agreement for the seven rounds of annotation. Alpha scores were calculated on a sentence basis (sentences split using Spacy).

Figure 3 shows the evolution of the mean Cohen's Kappa scores for each annotator. Mean kappa scores were obtained by calculating the scores for all pairs of annotators (considering just the "all" tag) and then calculating the mean value for each annotator. Values offer thus an indication of how much one annotator agrees with all the others. Two main trends are evident: first, there is a clear improvement through the seven rounds (moving from fair/moderate to substantial agreement); second, two annotators tend to always reach the highest scores, showing a better ability to agree with the others.

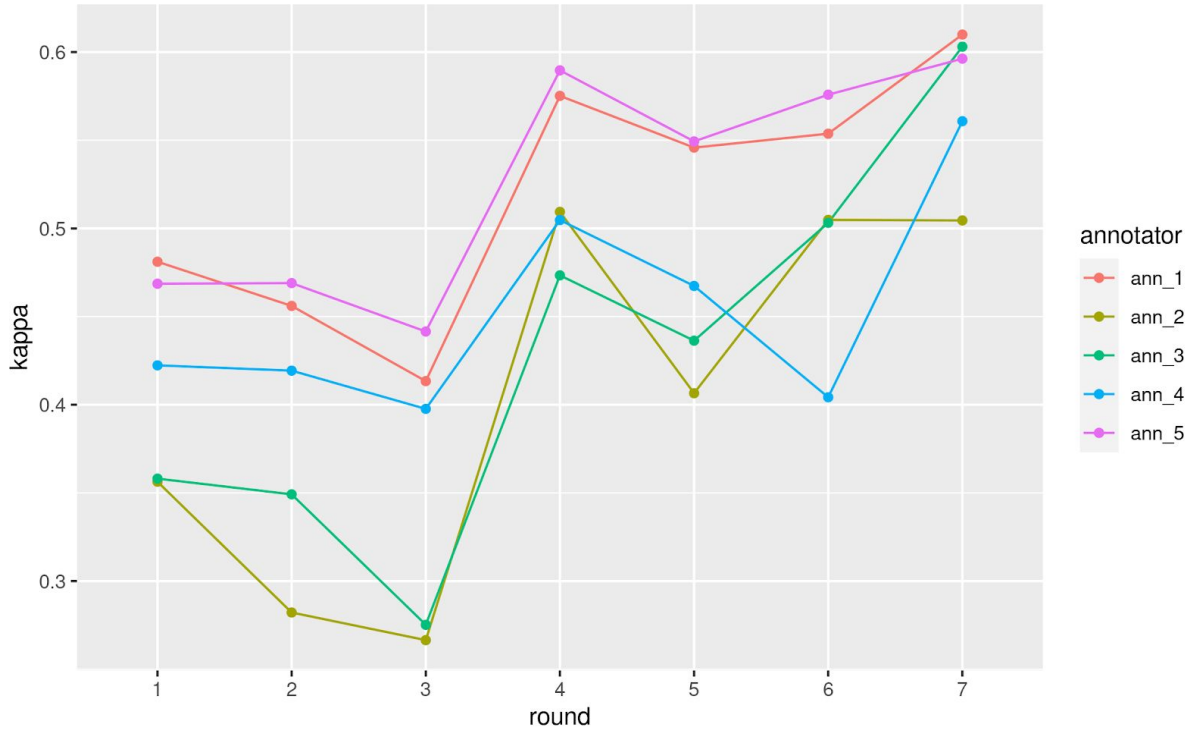

Figure 3. Inter-annotator agreement for the seven rounds of annotation. Kappa scores were calculated on a sentence basis (sentences split using Spacy).

Curation is currently in progress. However, first results confirm the already-observed trends. Mean agreement with the curator (mean Cohen's Kappa for the "all" tag) was 0.55 for the first round, while it reached 0.68 for the fourth.

3.2 Machine learning

We used several state of the art machine learning approaches to train a binary classifier on the annotated reviews, cf. [Lendvai et al. 2020]. When the current full dataset became available for training, a fine-tuned version of BERT [Devlin et al. 2018] reached 0.63 F-score on the target class, i.e., detecting absorption statements, and a linear regression model stacked on BERT predictions reached a mean average error of 0.08 (test set size: 149 reviews), cf. [Lendvai, Reichel, et al. 2020], which allow us to automate the annotation task and scale up the analysis of narrative absorption.

CONCLUSION

The "Mining Goodreads" project confirms the importance of interdisciplinary collaboration in the study of new phenomena such as digital social reading [Rebora et al. 2019]. The integration between empirical and computational methods also stimulates the definition of new research workflows in the wider context of digital humanities, where all the involved disciplines have the possibility to reach relevant goals: from the definition of a tool able to automatically recognize a complex linguistic and social phenomenon, to the improvement of the theoretical framework that defines it, to the broadening of literary studies towards unexplored grounds.

REFERENCES

Bálint, Katalin, Frank Hakemulder, Moniek M. Kuijpers, Miruna M. Doicaru, and Ed S. Tan. 2016. "Reconceptualizing Foregrounding: Identifying Response Strategies to Deviation in Absorbing Narratives." *Scientific Study of Literature* 6

(2): 176–207. https://doi.org/10.1075/ssol.6.2.02bal.
Devlin, Jacob, M.W. Chang, K. Lee, K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805.
Hakemulder, Jèmeljan, Moniek M. Kuijpers, and Ed S. Tan, eds. 2017. *Narrative Absorption*. Linguistic Approaches to Literature, volume 27. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
Hartung, Franziska, Peter Withers, Peter Hagoort, and Roel M. Willems. 2017. "When Fiction Is Just as Real as Fact: No Differences in Reading Behavior between Stories Believed to Be Based on True or Fictional Events." *Frontiers in Psychology* 8 (September). https://doi.org/10.3389/fpsyg.2017.01618.
http1. www.goodreads.com.
http2. www.goodreads.com/about/us.
Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. Association for Computational Linguistics. http://tubiblio.ulb.tu-darmstadt.de/106270/.
Kuijpers, Moniek M., Frank Hakemulder, Ed S. Tan, and Miruna M. Doicaru. 2014. "Exploring Absorbing Reading Experiences. Developing and Validating a Self-Report Scale to Measure Story World Absorption." *Scientific Study of Literature* 4 (1): 89–122.
Kuzmičová, Anežka, Anne Mangen, Hildegunn Støle, and Anne Charlotte Begnum. 2017. "Literature and Readers' Empathy: A Qualitative Text Manipulation Study." *Language and Literature: International Journal of Stylistics* 26 (2): 137–52. https://doi.org/10.1177/0963947017704729.
Lendvai, Piroska, Sándor Darányi, Christian Geng, Moniek Kuijpers, Oier Lopez de Lacalle, Jean-Christophe Mensonides, Simone Rebora, and Uwe Reichel. 2020. "Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation." In *Proceedings of The 12th Language Resources and Evaluation Conference*, 4835–4841. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.595.
Lendvai, Piroska, Uwe Reichel, Moniek Kuijpers, and Simone Rebora. 2020. "Ranking of Social Reading Reviews Based on Richness in Narrative Absorption." In *SwissText and Konvens 2020 5th SwissText & 16th KONVENS Joint Conference*.
Peer, Willie van, Jèmeljan Hakemulder, and Sonia Zyngier. 2012. *Scientific Methods for the Humanities*. Linguistic Approaches to Literature, v. 13. Amsterdam ; Philadelphia: John Benjamins Pub. Co.
Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J. Berenike Herrmann, Maria Kraxenberger, Moniek Kuijpers, et al. 2019. "Digital Humanities and Digital Social Reading." *OSF Preprint*, November. https://doi.org/10.31219/osf.io/mf4nj.
Rebora, Simone, Piroska Lendvai, and Moniek Kuijpers. 2018. "Reader Experience Labeling Automatized: Text Similarity Classification of User-Generated Book Reviews." In *EADH2018*. Galway: EADH. https://eadh2018.exordo.com/programme/presentation/90.
Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "BRAT: A Web-Based Tool for NLP-Assisted Text Annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–7. Association for Computational Linguistics.