jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

‘This book makes me happy and sad and I love it’

A Rule-based Model for Extracting Reading Impact from English Book Reviews

Marijn Koolen[1]
Julia Neugarten[2]
Peter Boot[2]

1. DHLab, KNAW Humanities Cluster, Amsterdam, Netherlands.
2. Literary Studies, Huygens ING, Amsterdam, Netherlands.

Citation
Marijn Koolen, Julia Neugarten, and Peter Boot (2022). "‘This book makes me happy and sad and I love it’. A Rule-based Model for Extracting Reading Impact from English Book Reviews". In: *Journal of Computational Literary Studies* 1 (1).
10.48694/jcls.104

Date published 2022-11-24
Date accepted 2022-03-31
Date received 2021-12-22

Keywords
reading impact, online book reviews, Goodreads, impact model

Reviewers
Katherine Bode
Peer Trilcke

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 1st Annual Conference of Computational Literary Studies 2022 at the Technical University of Darmstadt.

Abstract.
Being able to identify and analyse reading impact expressed in online book reviews allows us to investigate how people read books and how books affect their readers. In this paper we investigate the feasibility of creating an English translation of a rule-based reading impact model for reviews of Dutch fiction. We extend the model with additional rules and categories to measure reading impact in terms of positive and negative feeling, narrative and stylistic impact, humour, surprise, attention, and reflection. We created ground truth annotations to evaluate the model and found that the translated rules and new impact categories are effective in identifying certain types of reading impact expressed in English book reviews. However, for some types of impact the rules are inaccurate, and for most categories they are incomplete. Additional rules are needed to improve recall, which could potentially be enhanced by incorporating Machine Learning. At the same time, we conclude that some impact aspects are hard to extract with a rule-based model. When applying the model to a large set of reviews, lists of the top-scoring books in the impact categories show the model's prima-facie validity. Correlations among the categories include some that make sense and others that require further research. Overall, the evidence suggests that for investigating the impact of books, manually formulated rules are partially successful, and are probably best used in a hybrid approach.

1. Introduction

Online book reviews are an important source of data for analysing how people read books and how they describe reading experiences (Holur et al. 2021). This paper builds on our earlier work (Boot and Koolen 2020) in detecting the impact of reading fiction as it is expressed in online book reviews. That paper presented a rule-based model for measuring four categories of reading impact (affective, narrative, stylistic and reflective) in Dutch-language book reviews. As these rules are language-specific, the model cannot be used on the huge numbers of English-language reviews available online. In that article, we also mentioned potential types of reading impact that the model did not capture. In this paper, we present a model for measuring reading impact expressed in

English-language book reviews. We created this model by translating the Dutch model and adding rules for four new categories of impact: attention, humour, surprise and negative impact. To account for these new categories and refine the Dutch model, we re-categorised some rules and added more rules based on manual analysis of modes of expression in a corpus of Goodreads reviews. We analyse and validate the English model using crowdsourced ground truth annotations.

We formulate two research questions:

1. How effective is our adaptation of the Dutch model?
    (a) Can the new impact categories we add to the model be captured in a rule-based model? Can these new categories be meaningfully identified by human annotators?
    (b) Is adapting an existing rule-based model for use in another language a productive approach? Is our method of translating and changing rules an effective way to do this? What are the challenges and advantages of transferring knowledge or tools from Dutch to English through translation and adaptation?
2. Is a rule-based model a productive tool for assessing the impact of fiction as expressed in online book reviews? What are the advantages of a rule-based model compared to other approaches, such as Machine Learning (ML)?

First, we discuss the impact model and explain our selection of new impact categories in Section 2. Then, we describe how we created the rules that make up the English-language model by adapting the Dutch model in Section 3. We evaluate these adapted rules using the ground truth annotations and conduct an error analysis in Section 4. Human annotators recognise and distinguish categories of impact with some consistency, resulting in acceptable Inter-Rater Agreement. For several impact categories the rule-based model attains good performance in terms of precision and recall, but more ground truth data is needed to reliably validate some other categories, and for most categories more rules would be needed to cover the various ways impact can be expressed. In Section 5, we assess the quality of our results, using the model to detect reading impact in a large set of Goodreads reviews for a set of popular novels. We observe, aggregated over many reviews per novel, that the results mostly meet expectations. Finally, in Section 6, we formulate some suggestions for how to improve the model, and argue that taking a rule-based approach to assessing reading impact is a productive approach that may, in future work, be supplemented with other methods and tools.

Both the annotations and the rule-set used in the current paper are publicly available.

2. Impact Model and New Categories

2.1 Book Reviews as Data

Online book reviews are increasingly used to gauge reader response to books (Rebora et al. 2019; Spiteri and Pecoskie 2016). However, using online reviews for this purpose has its drawbacks and limitations: online reviews are not necessarily representative of all readers, they do not necessarily reflect readers' 'true' opinions, writing a review may influence the reading experience (Kuhn 2015) and they may be fabricated. We discuss these issues briefly in Boot and Koolen (2020). Prompted by the epistemological issues raised by one of this paper's reviewers, we want to make clear that we do not argue here that online reviews directly reflect the reading experience or necessarily provide insight into the general public's reading experiences. What the reviews can do is to provide us with insight into differences that exist among reviews or among reviewers. We hypothesise that such differences in the impact reported in reviews reflect differences in the experienced impact. By examining these differences, we can increase our understanding of how reading affects readers more generally. In this way studying online reviews complements existing research on reader response that uses interviews (Ross 1999; G. Sabine and P. Sabine 1983) or questionnaire data from readers reading selected short stories and passages in a controlled setting (Koopman 2016; Koopman and Hakemulder 2015; Miall and Kuiken 2002; Nell 1988). As online reviews are a more public form of reader response than interviews and questionnaires, we should remain aware that differences between reviews may also be attributable to social factors, such as the interactive nature of the reviewing platform or reviewers' desire to cultivate a persona or gain followers.

Using online book reviews as data also has some important advantages: the texts are accessible online in a digital format and they are primarily produced by groups of readers overlooked in much traditional literary scholarship. The writers of reviews on platforms like Goodreads are around 75% female (Thelwall and Kousha 2017), and users of Goodreads represent various nationalities and ethnicities (Champagne 2020). Thus, these reviews offer diverse perspectives that much of the field of literary studies lacks. We therefore consider them a useful source of information for literary scholarship in general and reception studies in particular.

Given the brevity of most online book reviews, we do not expect our model to perfectly identify all impact expressed in individual reviews. Instead, our aim is to develop a model that can identify relationships between aggregates of reviews grouped together by features like length, book genre or author gender, and the kinds of reading experiences described in reviews. In other words, we are producing a tool that enables literary scholars to assess the impact of books or collections of books on groups of readers by comparatively analysing the ways these books are reviewed online. Even though the representation of reading experience in reviews is nowhere near exhaustive, differences between these representations can nonetheless lead to insights into the impact of reading on reviewers. Eventually, we hope to be able to answer questions such as: How does the impact of the *Harry Potter* books change over the course of the series? How do readers differ in their responses, for instance by age, gender, or reading preferences? What patterns can we discern in the impact of specific genres or authors? Do reviewers review books differently depending on author gender or book popularity? Are there discernible patterns in how reviewers develop as readers?

We define impact as any effect a book has on its reader, large or small, permanent or fleeting. Thus, we conceptualise impact as a wider category than the personal or social longer-term benefits researched by e.g. Belfiore and Bennett (2007) or Usherwood and Toyne (2002). Impact includes what happens in the mind of the reader in response to reading fiction. As we did in Boot and Koolen (2020), where we based ourselves on Koopman and Hakemulder (2015), we investigate the following four categories of impact: *Reflection*, *Positive affect* and its two subcategories *Narrative feeling* and *Stylistic feeling*, and add four new categories.

2.2 Existing and New Impact Categories

In the final section of Boot and Koolen (2020) we express the expectation that smaller and more clearly defined impact categories might be better suited for validation in a survey. We therefore added four categories to our English version of the model: *Humour* as an additional subcategory of *Positive affect*, and three independent categories: *Attention*, *Surprise* and *Negative feeling*. This section introduces each new category and explains our motivation for adding it to the model.

Attention is one of the dimensions of Story World Absorption (M. M. Kuijpers et al. 2014), defined as "a deep concentration of the reader that feels effortless to them. As a consequence the reader can lose awareness of themselves, their surroundings and the elapse of time." Green and Brock (2000, p. 702) hypothesise that this feeling of absorption relates to changing beliefs and attitudes in readers. We chose *attention* as a category rather than suspense, although we consider the two closely related, because textual manifestations of attention can be distinguished more clearly than those of suspense. *Attention* is predicted in our model by terms such as 'immersed', 'absorbed' and 'engrossed.'

Humour, perceiving events or language as humorous, is a distinctive form of appreciation, related to but separate from stylistic or narrative feeling. Defining it as a separate category might make the categories of stylistic and narrative feeling more homogeneous. Humour is also relevant to the study of reading impact for its role in introducing young people to reading (Shannon 1993).

We added *Negative feeling*, which includes responses such as being bored or disappointed by a book, to help differentiate between positive and negative expressions of impact. Although some research examines the negative effects of reading (Schmitt-Matzen 2020) and a negative response to prescribed reading (Poletti et al. 2016), previous research has overwhelmingly focused on trying to validate the hypothesis that reading is good for personal development and social behaviour (Koopman and Hakemulder 2015), while negative feelings towards reading are often overlooked. Rather than having a single *Negative feeling* category, we could have distinguished positive and negative feeling towards narrative, style and humor. We decided against this option because we expected (based on exploratory analysis of reviews using (Kilgarriff et al. 2014)) that negative stylistic feeling and negative references to humor would not occur frequently enough to be detectable.

We included *Surprise* as a category of impact, because it shows engagement with a story. If surprises in reading are defined unexpected story elements, experiencing surprise requires one to have expectations of a book which are subsequently defied, and these expectations are a sign of engagement. We therefore considered including *Surprise* in *Narrative feeling*. On the other hand, surprise also indicates cognitive processing (Tobin 2018) and could thus be considered part of *Reflection*. It is also possible to conceptualise *Surprise*, which can incorporate elements of "violence and enlightenment, physical attack and aesthetic pleasure" (Miller 2015) as a type of *Stylistic feeling*. In the end, to navigate this complex process of conceptualisation, we chose to measure *Surprise* by itself. Correlations with other categories could help us theorise the nature of *Surprise* further.

2.3 Definitions

These considerations led to the following definitions for eight categories of impact:

- **Attention**: the reader's feeling of concentration or focus on their reading.
- **Positive affect**: any positive emotional response to the book during or after reading. A feeling is positive if it contributes to a positive reading experience, so even sad or awful story-events can contribute to a positive affective response.
    - **Narrative feeling**: a subcategory of positive affect, specifically response to a book's narrative properties, including feelings about storylines, characters, scenes or elements of the story world.
    - **Stylistic feeling**: a subcategory of positive affect, specifically response to a text's stylistic properties such as feelings of admiration or defamiliarisation about its tone, choice of words, use of metaphor or the way the sentences flow.
    - **Humour**: a subcategory of positive affect, specifically a response of laughter, smiling or amusement; the effect of any type of humour in the text.
- **Surprise**: a feeling of surprise at some element of the book, such as a plot development, part of the story world or a stylistic feature.
- **Negative feeling**: feelings of dislike or disapproval towards any element of the book. This could mean a dislike for a storyline or character or a feeling of boredom or frustration with the book as a whole. A feeling is negative if it contributes to a negative reading experience, so unsympathetic characters or dark story elements that a reviewer appreciates as part of a story do not fit within this category.
- **Reflection**: any response to a reading experience that makes the reader reflect on something from the book, such as a theme or topic, or on something in the real world.

3. Methods

This section introduces how the impact model works and explains the method of its validation.

3.1 Model Development

Our model uses a set of rules to identify different types of impact expressed in individual sentences of reviews, similar to the setup used in Boot and Koolen (2020). Each rule belongs to a category and consist of an impact term, an impact term type and in some cases a condition. For each combination of sentence and rule the software checks whether the impact term is present in the sentence and, if there is a condition, whether that condition is met. If so, it outputs a rule match with the associated impact type.

Impact terms can be lemmas or phrases. If they are lemmas, their impact term types include a POS-tag. For example, if the impact term is 'mesmerize', and the type is 'verb' the software will check for each word in the sentence whether it is a verb form with that lemma. POS-tags can also be 'noun', 'adjective' or 'other'. In phrasal impact terms, no lemma or POS information is used, and terms can contain wildcards (*), so 'redeeming qualit*' finds both 'redeeming quality' and 'redeeming qualities'. Phrases consist of groups that are matched to tokens in the input sentences. A group can be a single word or a set of alternatives, such as '(hard|difficult)'. A phrase can be continuous or discontinuous. In a continuous phrase the groups must match a set of contiguous tokens. In a discontinuous phrase each group must match a token in the same sentence in the same order as in the phrase, but they need not be adjacent. For examples, see Table 1.

Impact			**Condition**	
type	***term***	***term type***	***aspect***	***negate***
Attention	on the edge of (my|your) seat	phrase-continuous	–	–
Positive affect	makes (me|you|reader) sad	phrase-discontinuous	–	–
Narrative feeling	enamoured	lemma-adj	reader	–
Stylistic feeling	elegant	lemma-adj	–	–

Table 1: Example rules from the English reading impact model.

Conditions can also have different types. Most common is a reference to one of six groups of book aspect terms: *plot*, *character*, *style*, *topic*, *reader* and *general*. For example, aspect terms in the *reader* group are words referring to the reader, such as 'I', 'you', 'the reader' and the *general* group includes words like 'book' and 'novel'. The implied condition is that one of the words from the aspect group must occur in the same sentence as the impact term. Thus, a rule linking the impact term 'great' to the aspect group *style* results in a hit when the word 'great' is present in combination with 'writing', 'language', 'prose' or other words in the *style* category. Conditions can also be groups of individually named words, such as '(part|series|sequel)'. It is also possible to negate a condition. In that case, the impact term may not be combined with words from the condition. For example: 'engage' is an impact term related to *Narrative feeling*, unless it is combined with 'to' because 'engaged to' is more likely to refer to marriage than to narration.

To create the rules, we began by translating the 275 rules of Boot and Koolen (2020). To account for the new impact categories, we reassigned some rules to different categories. We also created new rules by manually examining a large collection of Goodreads reviews to find terms related to impact that online reviewers use. In total, the English model has 1427 rules. The growth of the set of rules has three main reasons. Firstly, the addition of four categories required adding many rules. Secondly, there are many possible translations or equivalents for the words and expressions used in the Dutch model. For example, some words relating to emotional investment in the Dutch model led to eight new rules in the English model containing various verbs combined with the noun 'heart' ('break', 'steal', 'touch', 'rip' and others). Thirdly, there are many more reviewers writing in English. Many have their own national or regional variety of English and many of them are not native speakers. Because of this variety, we assumed that the range of expressions used in reviews would be larger than in Dutch. To account for this range, we added many idioms based on manual analysis of a corpus of reviews from Goodreads. As we found (Boot and Koolen 2020) that human annotators often detected impact that their Dutch impact model overlooked, we expect that adding more rules will lead to a better model.

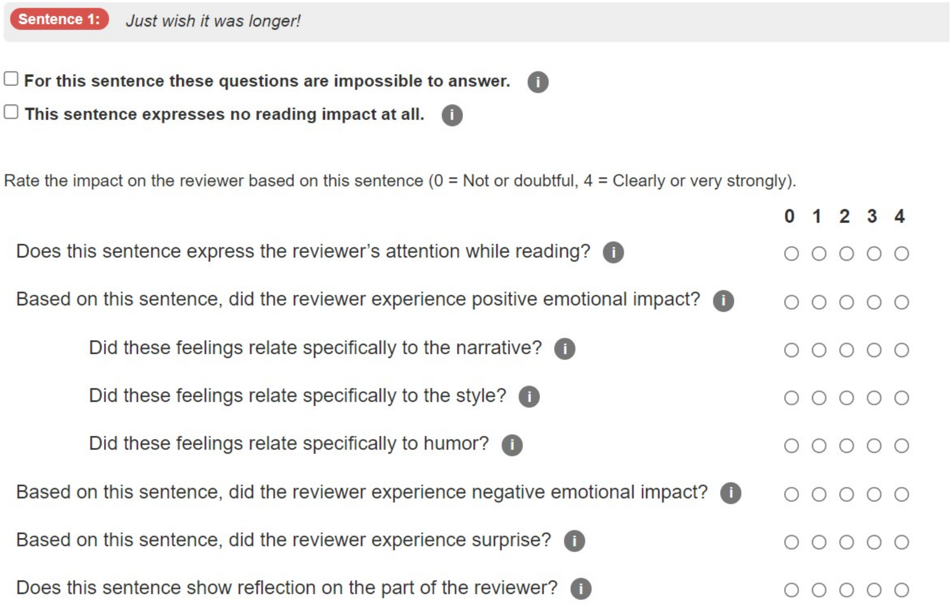
Our choice to follow the rule-based approach needs be compared to alternative approaches, such as creating ground truth annotations and using ML techniques to train a generalised model. Our main reason to use rules instead of ML is that we expect ML to require many more ground truth annotations to train and test a stable and effective model that can capture subtle expressions of impact. Our model was developed ahead of gathering ground truth annotations to evaluate it (as discussed in the next sections). An ML model only learns from the annotated examples, while our rules potentially also cover cases not seen in the ground truth. If the evaluation shows that our model captures the different impact categories well, then we have reason to assume that the rule generation process achieved its aim and that the approach generalises well. With ML this is not necessarily so, although the recent advances with context-sensitive token-based word embeddings and fine-tuning of large pre-trained transformer models like BERT (Devlin et al. 2018) allow such approaches to better capture latent meanings (Ehrmanntraut et al. 2021; B. Wang et al. 2019; Yile Wang et al. 2019) and generalise beyond the surface forms of the annotated impact expressions. The manual rule-based approach can be combined with ML to either derive more rules or to directly assign sentences to impact categories. We discuss this further in Section 6.

3.2 Ground Truth Annotations

The rules we formulated determined how the impact model defines the various categories of impact. Next, we needed to verify that the rules we had formulated correctly operationalised the intended categories of impact. After all, the definitions implicitly created through the formulation of our impact rules might not agree with a commonsense idea of how these categories of impact are expressed. To validate our impact rules, we surveyed recipients of relevant mailing lists, students and conference attendees. We asked the participants to annotate sentences from reviews on the presence of the eight impact types. The sentences were sampled from a collection of 15 million English-language Goodreads reviews, crawled by Wan and McAuley (2018) and Wan et al. (2019), and parsed using spaCy.[1] We manually removed sentences that we considered impossible to annotate, such as sentences containing only punctuation or incorrectly split (partial) sentences. Each sentence was annotated by at least three different annotators. After reading an explanation, each annotator was presented with ten sentences to annotate. Each annotator could annotate as many sentences as they wanted. The questions were presented to them as shown in Figure 1.

Participants could rate the presence of all eight categories of impact on a five-point scale from 0 (not or doubtful) to 4 (clearly or strongly). They also could indicate that the

[1] The sentences were from a held-out set of reviews, not used to create the impact rules. We used spaCy version 2.3, https://spacy.io



Sentence 1: *Just wish it was longer!*

☐ **For this sentence these questions are impossible to answer.**

☐ **This sentence expresses no reading impact at all.**

Rate the impact on the reviewer based on this sentence (0 = Not or doubtful, 4 = Clearly or very strongly).

	0	1	2	3	4
Does this sentence express the reviewer's attention while reading?	○	○	○	○	○
Based on this sentence, did the reviewer experience positive emotional impact?	○	○	○	○	○
Did these feelings relate specifically to the narrative?	○	○	○	○	○
Did these feelings relate specifically to the style?	○	○	○	○	○
Did these feelings relate specifically to humor?	○	○	○	○	○
Based on this sentence, did the reviewer experience negative emotional impact?	○	○	○	○	○
Based on this sentence, did the reviewer experience surprise?	○	○	○	○	○
Does this sentence show reflection on the part of the reviewer?	○	○	○	○	○

Figure 1: Questions in the survey.

questions were impossible to answer, for example if the text only contained gibberish or required more context to interpret, or that a sentence expressed no reading impact at all, for example if it contained only a factual statement about a book. We ran the survey from October 2020 until April 2021.

4. Evaluation

In this section, we assess agreement among the annotators and between the annotators and our model, and analyse which impact categories our model can meaningfully identify.

The survey resulted in 266 sentences that were annotated by at least three annotators, with ratings by 79 different annotators. This number excludes sentences judged to be impossible to annotate. The majority of annotators rated 10 sentences, some stopped after only a few sentences, and others annotated multiples of 10 (up to 80). The distribution of ratings per impact type is shown in Figure 2. On the left, only the zero ratings are shown, meaning a rating to indicate that a particular impact type is not present at all. *Positive affect* has the fewest ratings of 0, with just over 40%, while *Stylistic feeling* and *Surprise* have around 70% 0 ratings and *Humour* has more than 80%. Over all categories, 69% of the ratings have score zero. On the right, the distribution of ratings 1–4 are shown, also with distinct differences between types. *Positive affect* and *Narrative feeling* tend to get high ratings (3 or 4), while *Attention* and *Surprise* get mostly low ratings (1 or 2).

4.1 Inter-Annotator Agreement

In Boot and Koolen 2020, we calculated inter-annotator agreement using the Inter-Rater Agreement (IRA) statistic $r^*_{wg} = 1 - \frac{S^2_X}{\sigma^2}$, where S^2_X is the variance of the ratings for a sentence and σ^2 is the expected variance based on a chosen theoretical null-distribution

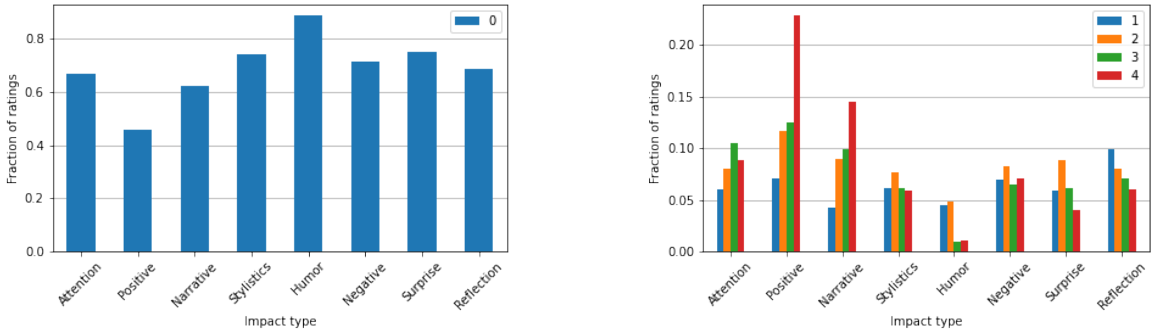


Figure 2: Fraction of 0-ratings among all ratings (left) and fraction of positive ratings (1, 2, 3 or 4) among all ratings (right).

Category	% all zero	r^*_{wg}	κ
Attention	0.37	0.58	0.27
Positive affect	0.26	0.71	0.57
Narrative	0.36	0.55	0.40
Style	0.49	0.72	0.29
Humor	0.72	0.91	0.19
Negative	0.56	0.79	0.52
Surprise	0.50	0.74	0.25
Reflection	0.39	0.60	0.19

Table 2: Inter-annotator agreement per impact category averaged over 266 sentences. Agreement measures are r^*_{wg} and Fleiss' Kappa.

(Lindell and Brandt 1997). We used the same r^*_{wg} measure, but with a uniform null-distribution instead of an inverse triangular one (which assumes annotators tend to pick ratings at the two extremes), given that we observe a more uniform distribution of positive ratings when combining ratings across all categories and a larger fraction of zero ratings (so the overall variance is closer to a uniform distribution than to an inverse triangular distribution). In addition, we report Fleiss' Kappa (κ) on binarised ratings where any rating above 0 is mapped to 1, as is more commonly reported in sentence annotation tasks for e.g. sentiment analysis (Alm and Sproat 2005; Schmidt et al. 2018; Sprugnoli et al. 2016). Finally, we also report the number of sentences rated zero on a particular impact category by all three annotators, to get insight into how commonly each impact category is observed.

Agreement is moderate (0.51-0.70) to very strong (0.91-1.00) according to r^*_{wg} (column three in Table 2), but the κ scores are much lower, in the range of 0.20 – 0.50 (column four). Scores in this range are common for related tasks like sentiment annotation (Alm and Sproat 2005; Klenner et al. 2020; Schmidt et al. 2018; Sprugnoli et al. 2016). The low κ of the more commonly observed categories should not be interpreted as low agreement, because in the original five-point scale the difference between 0 and 1 is small, while in the binarised version it is counted as disagreement.

To understand how the differences between r^*_{wg} and κ should be interpreted, we look at the number of sentences for which all three annotators agreed on a rating of zero. Since the majority of the ratings (see the left-hand side of Figure 2) is zero, this can easily lead to a high r^*_{wg}, especially for categories that are rarely rated above zero. If a category is rarely observed, it is easy for annotators to agree on the many sentences where it is clearly not present, but they might disagree on the few sentences where at least one annotators thinks it is present. Only 26% of all sentences are rated zero on *Positive affect* by all three annotators, so its high r^*_{wg} is not caused by being rarely observed. In contrast, for *Humour*, 72% of the sentences are rated zero by all annotators, meaning it is rarely observed. For this category, a high r^*_{wg} could be caused by agreement that the category is rare, thus masking disagreement on which sentences do express impact of humour. The κ score of 0.19 (below the conventional 0.2 threshold for weak agreement) signals that agreement is lacking. For *Reflection*, only 39% of sentences are rated zero by all annotators, so this category is not uncommon, but the κ score of 0.19 also suggests a lack of agreement. We stress again that a low κ does not necessarily mean lack of agreement, as the binarisation removes information from the five-point rating scale, but for *Humour* and *Reflection* these combined measures strongly suggest either that these categories are difficult to identify with our current definitions, or that reliable annotation of these categories requires more training than of the other categories.

The disagreement among annotators signals that this task is difficult and that some types of impact are more subjectively interpreted than others. This could indicate that we need to discard the categories with really low agreement. However, several recent papers suggest that disagreement between annotators is not necessarily a problem and should not be removed from the published annotation dataset (e.g. Gordon et al. 2021), but should either be retained in the form of an opinion distribution (Basile 2020; Klenner et al. 2020) or a special class label *Complicated* (Kenyon-Dean et al. 2018). Since our data is based on a rating scale, it makes sense to distribute the annotated sentence data with the full rating distributions. In the following sections, we discuss whether all impact categories should be retained in the ground truth data and the rule-based model.

4.2 Evaluating the Model

To compare our model against the ratings of the human annotators, we select the median of the three ratings per sentence and impact category as the ground truth rating and compare that to whether our model finds at least one matching impact rule for that category in the sentence. If the model works well, then it should find matching rules for an impact category in sentences that received a high median rating from human annotators.

We measure recall, precision and F_1 of our model's performance on the annotated sentences, using two different binarisations. As we have a 5-point rating scale, we want to know if our model finds impact in sentences that clearly express impact, that is, where the median rating is high, i.e. 3 or 4, but also for sentences that express any impact at all, i.e. those with ratings of 1 or higher. The results are shown in Table 3, with the number of sentences that have a binary rating of 1 for each binarisation (columns 3 and 6). The model scores above 0.7 precision on five of the eight categories for binarisation $r_{median} \geq 1$: *Attention*, *Positive affect*, *Narrative feeling*, *Humour* and *Negative feeling*. In the majority of cases, the matching rules for these aspects correspond to the type of impact identified by the median annotator, and therefore at least two of the three annotators. For *Stylistic feeling* and *Reflection* it scores around 0.5 precision, so the matching rules incorrectly signal impact in half of the cases. For *Surprise* the model

	Model	$r_{median} \geq 1$				$r_{median} \geq 3$			
Impact	# Sent.	# Sent.	Prec.	Rec.	F_1	# Sent.	Prec.	Rec.	F_1
Attention	9	83	0.78	0.08	0.15	44	0.78	0.16	0.26
Positive	90	148	0.82	0.50	0.62	102	0.59	0.52	0.55
Narrative	39	101	0.72	0.28	0.40	60	0.51	0.33	0.40
Stylistic	8	59	0.50	0.07	0.12	26	0.50	0.15	0.24
Humour	7	18	1.00	0.39	0.56	4	0.57	1.00	0.73
Negative	15	68	0.73	0.16	0.27	40	0.60	0.23	0.33
Surprise	2	51	0.00	0.00	0.00	17	0.00	0.00	0.00
Reflection	19	68	0.53	0.15	0.23	19	0.26	0.26	0.26

Table 3: Model evaluation per impact category on 266 sentences, with number of sentences for which the model identifies impact (column 2), and precision and recall of our model for binarisation of ratings based on median rating $r_{median} \geq 1$ and $r_{median} \geq 3$
.

completely fails. It only finds *Surprise* in two sentences—both of which are incorrect according to the ground truth—while there are 51 sentences with a median rating of at least 1. For binarisation $r_{median} \geq 3$, precision is mostly lower, showing that the model regularly predicts impact where human annotators consider it doubtful. *Humour* is rarely observed by the annotators, with low agreement, and our model also rarely finds matching rules, but with high precision for $r_{median} \geq 1$ and high recall for $r_{median} \geq 3$. When annotators agree that *Humour* is clearly expressed, our model detects it (in the few cases in this ground truth dataset), and when our model detects *Humour*, it is in places where annotators perceive *Humour* to some extent. Two examples where annotators and our model clearly agree demonstrate this. For the sentence "I loved Blaire's personality she was sassy, funny, extremely witty, I laughed out loud frequently, much to my embarrassment." our model has three matching rules,*funny*, *witty* and *laugh out loud*, and the annotators gave an average rating of 3. For the sentence "We actually bought a copy for our music history teacher who would appreciate the humour in this book (he was Jewish, sarcastic, clever)." our model has two matching rules, *humour* and *sarcastic* (which in this sentence does not refer to impact of the book) and annotators gave an average rating of 3.33. This sheds further light on the low Fleiss' Kappa scores for *Humour*. There are clear cases where annotators agree that humour is expressed, so the low agreement seems to come from doubtful cases where some annotators are not sure and give a low rating of 1 or 2 and others say it is not expressed. The binarisation we used to compute Fleiss' Kappa creates a complete disagreement in such doubtful cases, where the original five-point ratings signal only slight disagreement. The model performance suggests that, although we need more ground truth annotations and perhaps a better definition to improve agreement, this is a viable category to include.

The generally low recall scores show that our model misses many expressions of reading impact. This suggests that our set of impact rules is incomplete. It could be fruitful to add more rules to our impact model. However, if expressions of impact in reviews have a long tail distribution, the dwindling number of hits for each added rule may not be worth the time and labour it takes to formulate, add, and validate additional rules. One interpretation of the precision and recall scores is that our approach of translating and extending our Dutch impact model is viable for most of the categories. With additional rules and some improvements to the existing rules, the model could capture enough of the expressed reading impact in individual reviews to derive a reliable overall estimate of a book’s impact, at least for books with more than a handful of reviews. The only impact types where the model fails are *Surprise*, *Stylistic feeling* and *Reflection* where the model not only misses many expressions of impact, but also makes many mistakes. However, it is also possible that some impact is expressed across multiple sentences, in which case additional rules would have to operate on larger text units, which makes them harder to devise and more sensitive to errors. Therefore, another interpretation of the results is that it is more effective to combine our rule-based approach with a ML process that learns to assign impact categories directly to individual sentences or to whole paragraphs or reviews.

4.3 Error Analysis

Annotators found impact in many instances where the model failed to detect it. For example, the model scored a 0 in the positive emotion category for the sentence “I was born to love this book,” which received a rating of 4 from all annotators. This suggests we should add rules to increase the sensitivity of the model. We should also revise the way that the model processes impact terms and negations to nuance results. Currently, the model marks the presence of negative terms like ‘skim’ as negative impact, but it turns out that this is not always accurate: “I didn’t skim at all” actually indicates positive impact. The negation of the negative impact term ‘skim’ should flip the predicted impact to positive. To improve performance on sentences with such negations of typical sentiment words, we could adopt the sentiment flipping technique used in the VADER sentiment analyser (Hutto and Gilbert 2014). This technique looks for negations in the word tri-gram preceding a sentiment term, which captures almost 90% of the negated sentiments in their ground truth data. However, negation should not always flip the valence from positive to negative or vice versa (Dadvar et al. 2011; Socher et al. 2013). When a reviewer says that a book is ‘not terrible’ they probably don’t mean to say it is good.

The responses to the survey showed that annotators struggled to understand some categories and regularly disagreed over them, albeit to a different degree for different categories. For instance, the sentence “And then there was Jacob O’Connor,” which we feel expresses no impact, was rated by annotators with a score of 3.5 in the surprise-category. Annotators also found *Attention* difficult to distinguish from *Positive affect* and *Narrative feeling*. They struggled with negative story elements that may contribute to a positive reading experience, such as a ‘creepy’ character. Respondents tend to annotate such sentences as negative impact, while that is often impossible to judge without context. In another example, annotators judged the sentence “My soul is beautifully crushed” to indicate negative impact, but we think that a reviewer who writes this is expressing positive impact. These differences between annotator ratings and our own conceptions of impact categories point towards one of the complexities of developing computational models for literary studies: while defining categories of impact and formulating rules for our model, our own subjective understanding and academic knowledge of impact categories and the impact of reading became part of the model we produced. These conceptions may not necessarily align with the conceptions of other people. To resolve issues of annotator agreement, we could consider recruiting annotators with a background in reception studies or literary studies for future research, since they will presumably have a shared understanding of these impact categories based on the scholarly literature. Therefore, these annotators would probably be better equipped to distinguish and detect our eight impact categories, but it is also possible that they would skew results with their pre-existing definitions of the categories. Another option would be resolving disagreement between annotators using the method for interrater disagreement resolution outlined by Oortwijn et al. (2021), or recruiting annotators from within the community of people writing English-language reviews on Goodreads. This way, we could validate the model using conceptions from within the community we are studying. While we tried to do this by contacting the moderators of various Goodreads groups, we received little response.

In the end, developing a flawless model to measure how reading impact is expressed in online reviews may be impossible, because of the subjectivity and fluidity of the categories such a model tries to measure. In the act of operationalising impact categories through rulesets, some of their polysemic meanings are inevitably lost. Nonetheless, we believe that our current imperfect model has pointed us towards some interesting insights into the impact of reading expressed in our corpus of reviews. We discuss these insights in Section 5.

5. Analysing Reading Impact of Fiction

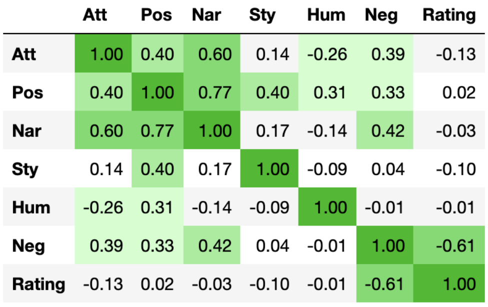
In this section, we analyse the impact identified by our model by applying it to a collection of 1,313,863 reviews of 402 well-known books, from the Goodreads crawl introduced in Section 3.2. As the results from the previous section cast doubt on the viability of measuring some of the categories of impact, this section disregards *Surprise* and *Reflection*. We selected books with at least 10 reviews in both Dutch and English so that, in future research, we may compare the current and future versions of the English-language model against the Dutch model.

5.1 Impactful Books

Our model generated a rating for each of the 402 books in each of the model's categories. This rating gives an indication of how often a specific type of impact was mentioned in a specific review. After normalising the scores for the length of the reviews we computed which books scored highest and lowest in each category. Table 4 lists the books scoring highest on *Stylistic feeling* and *Humour*. The left column contains mostly literary classics that received high critical acclaim; we would expect those novels to score high on *Stylistic feeling*. The right column contains mostly books that are well-known for their comic appeal. Similar lists for other categories are not always easy to evaluate, for example because lesser-known novels appear in the list or because there is no canon of novels that evoke a high level of narrative engagement, the way there is one, albeit a fuzzy and contested one, for literary novels. Still, the results suggest that our rules are detecting some impact accurately. For example, we expect that non-fiction titles would score low on *Narrative feeling*, and we find that the four worst-performing titles in terms of *Narrative feeling* are non-fiction titles, including Marie Kondo's *The Life-Changing Magic of Tidying Up*. These results provide prima facie evidence for the validity of the rules that we use to define these impact categories.

	Title	Author
Stylistic feeling	Monsieur Linh and his Child	Philippe Claudel
	Lolita	Vladimir Nabokov
	All the Light We Cannot See	Anthony Doerr
	Stoner	John Williams
	The Sense of an Ending	Julian Barnes
	The Discovery of Heaven	Harry Mulisch
	HHhH	Laurent Binet
	The Vanishing	Tim Krabbe
	A Visit from the Goon Squad	Jennifer Egan
	The Book Thief	Markus Zusak
Humour	Weird Things Customers Say in Bookshops	Jen Campbell
	Look Who's Back	Timur Vermes
	The Hitchhiker's Guide to the Galaxy	Douglas Adams
	The Secret Diary of Hendrik Groen, 83¼ Years Old	Hendrik Groen
	The Hundred-Year-Old Man Who Climbed Out of […]	Jonas Jonasson
	The Girl Who Saved the King of Sweden	Jonas Jonasson
	Me and Earl and the Dying Girl	Jesse Andrews
	The Rosie Project	Graeme Simsion
	A Totally Awkward Love Story	Tom Ellen
	Geek Girl	Holly Smale

Table 4: Top ten titles on stylistic feeling and humour



	Att	Pos	Nar	Sty	Hum	Neg	Rating
Att	1.00	0.40	0.60	0.14	-0.26	0.39	-0.13
Pos	0.40	1.00	0.77	0.40	0.31	0.33	0.02
Nar	0.60	0.77	1.00	0.17	-0.14	0.42	-0.03
Sty	0.14	0.40	0.17	1.00	-0.09	0.04	-0.10
Hum	-0.26	0.31	-0.14	-0.09	1.00	-0.01	-0.01
Neg	0.39	0.33	0.42	0.04	-0.01	1.00	-0.61
Rating	-0.13	0.02	-0.03	-0.10	-0.01	-0.61	1.00

Figure 3: Pearson correlation coefficients between impact types and rating of reviews aggregated per novel.

5.2 Correlations between Impact Types

In this section, we analyse the correlation between impact types and the correlation between selected impact types and the average rating of reviews, when aggregated per novel, for the same set of 402 novels. For this analysis, we computed impact score per category based on the recommendation of Koolen et al. (2020), where we suggest weighing the number of impact rule matches per review by the log-length of the review in number of words. This weighing should account for the fact that long reviews potentially have more impact matches without actually indicating stronger impact. The Pearson correlations are shown in Figure 3, with levels of correlation above 0.2 highlighted in green. Unsurprisingly, *Positive affect* is positively correlated with its components *Narrative feeling*, *Stylistic feeling* and *Humour*. We discuss the correlations of *Attention* and *Negative feeling* with the other impact categories and these categories' correlations with reviewer rating.

5.2.1 Correlations of *Attention*

The most important correlation (.60) for *Attention* is with *Narrative feeling*. This suggests that *Narrative feeling* draws readers in and leads to a sense of absorption and immersion. Attention-related questions are also an important part of the Story World Absorption Scale (M. M. Kuijpers et al. 2014). On the other hand, the lack of correlation between *Attention* and *Stylistic feeling* suggests that stylistic appreciation is not central to absorption. *Attention* is weakly negatively correlated with *Humour*. In an analysis of evaluative terms, Knoop et al. (2016) distinguish between emotionally charged terms such as 'sad' and 'beautiful' and more cognitive terms such as 'funny' or 'humorous'. The relationship between cognitive and emotional impact is an area of further research that could help refine our model and generate insight into different ways that readers evaluate texts and texts impact readers.

5.2.2 Correlations of *Negative feeling*

It is surprising that *Negative feeling* is weakly to moderately positively correlated with *Attention*, *Positive affect* and *Narrative feeling*. As this is not just a book-level effect (*Positive* and *Negative feeling* are also correlated within individual reviews), we speculate that these correlations occur because negative terms are often used concessively, as in 'the plot may be a bit unrealistic but the characters are lovely'. More research on these correlations is needed.

5.2.3 Correlations of Impact Categories and Reviewer Rating

On Goodreads, reviewers have the option of rating a book on a five-star scale in addition to, or instead of, providing a written review. Only one impact category shows a correlation with reviewer rating: *Negative feeling*. The moderate negative correlation suggests that negative terms are not just used concessively but often do express a lack of appreciation.

The lack of correlation between rating and the other impact categories may at first seem surprising. Positive feeling, as measured by sentiment analysis tools, is known to predict rating (De Smedt and Daelemans 2012). We would also expect *Attention*, which is closely related to enjoyment (M. M. Kuijpers et al. 2014), to correlate positively with rating. This lack of correlations could indicate that the impact model succeeds in extracting new information, independent from rating, from the review text. In other words, reviews may be more than just a textual representation of the associated rating.

The correlations among impact types, or lack thereof, as well as those between impact types and rating, call for further analysis of the nature of their relation. For instance, the collection of reviews is skewed towards high ratings (over 71% of reviews have a positive rating of 4 or 5 stars, while less than 10% have a negative rating of 1 or 2 stars). As there is a lot of variation in positive impact matches in the positive reviews, the positive reviews dominate at any level of positive impact, leading to a low correlation. Moreover, reviews also have high variance in terms of length, with long reviews having a wider range of possible values for positive impact or any other type of impact than short reviews. If we bin reviews on the logarithm of their length and look at correlations between rating and impact within each bin, we get somewhat higher positive correlations. For reviews of more than 20 words, reviews of roughly the same length show a correlation of 0.13-0.16 between rating and positive impact. This is still no more than a weak correlation, given further evidence that reviews reflect more than just a rating-related evaluation.

Reader characteristics may also influence the relation between review and rating. For instance, we found a negative correlation between impact in the *Reflection* category and reviewer ratings (not shown in Figure 3). This could mean that reviewers are less appreciative of books that encourage reflection. But it could also mean that readers who engage in more reflection generally give more moderate ratings. More generally, this prompts the question how rating behaviour relates to reading preference and other reader characteristics.

6. Discussion

Our findings allow us to address our two main research questions and to indicate a number of areas for future research into the impact of fiction and the usefulness of computationally measuring and analysing that impact in online book reviews. In the upcoming research project *Impact and Fiction* we will build on the findings presented in the current paper.

6.1 Conclusions

1. *How effective is our adaptation of the Dutch model?*
   Based on the results from the English impact model so far, the model is effective in some categories but not all of them. For several impact categories the rule-based model attains good performance in terms of precision and recall, but more ground truth data is needed to reliably validate some other categories, and for some categories more rules are needed to cover the various ways impact can be expressed. When ranking books by scores in individual impact categories, the model appears to do a good job. In future work, we intend to compare the English impact model presented in this paper with the existing Dutch model.

   (a) *Can the new impact categories we add to the model be captured in a rule-based model? Can these new categories be meaningfully identified by human annotators?*
   We added four new impact categories to the impact model described in Boot and Koolen (2020), in the hope that adding more categories would lead to a more fine-grained and accurate model. Some of these newly added categories proved difficult for annotators to identify consistently. For example, annotators frequently seemed to confuse *Attention* and *Narrative feeling*. For example, according to the annotators "Lots of twists and turns and good characters" indicated *Attention* as well as *Narrative feeling*. Yet we, and the rules of our model, see this sentence as indicating only *Narrative feeling*. Conversely, annotators labelled the sentence "The third book of the trilogy is just as compelling as the other two" as both *Narrative feeling* and *Attention*, while our model and conceptualisation of categories would only see it as *Attention*. Such overlap, disagreement or confusion between categories shows that, similar to the original categories, identifying which sentences express a specific type of impact remains a difficult and subjective task.

One way of approaching this issue might be to compare the correlations between the impact categories as established by our model and those between the impact categories as rated by the annotators. That could provide us with a sense of how the annotators' conceptualisation of the impact categories differs from our model's conceptualisation. All the same, the combination of inter-annotator agreement analysis, evaluation of the model based on the ground truth annotations, and comparing the reading impact of novels identified in sets of reviews, already illustrates that some of the new impact categories can be meaningfully identified using a rule-based model.

(b) *Is adapting an existing rule-based model for use in another language a productive approach? Is our method of translating and changing rules an effective way to do this? What are the challenges and advantages of transferring knowledge or tools from Dutch to English through translation and adaptation?*
Our results indicate that the translation of the rules, in combination with adding new rules specific to English, is a viable approach to building a reading impact model for English-language reviews and expanding on the existing Dutch model. However, since human annotators detect impact in many words and phrases that the model disregards, the model is certainly not complete. Also, adapting the model was a labour-intensive process.

2. *Is a rule-based model a productive tool for assessing the impact of fiction as expressed in online book reviews? What are the advantages of a rule-based model compared to other approaches, such as ML?*
A rule-based model has advantages and drawbacks compared to other approaches, like ML. Rule-based approaches are more transparent than trained ML models because users can inspect every rule and understand how the model arrived at a specific decision. With ML models, especially neural network-based models, the knowledge is distributed over and represented by a large number of weights between the network nodes. In a rule-based model researchers can add or translate impact-rules, and this way they can adapt the tool to specific research questions and language domains without requiring large amounts of ground truth annotations. Moreover, for fine-grained annotation in specific domains, like identifying expressions of different types of reading impact, it can be difficult to attain good performance with ML, as ML models need to be trained on domain-specific data to adapt to the domain-specific terminology and nuances (Mishev et al. 2020; Thelwall et al. 2010; Wu and Huang 2016), which requires large amounts of training data. For instance, for the simpler task of sentiment polarity classification, many thousands or tens of thousands of annotated examples are needed (Mishev et al. 2020; Yao and Yan Wang 2020). At the same time, formulating and validating rules is also a labour-intensive process and our model did not attain great results for every category. However, the impact model presented in this paper could potentially be used to gather relevant data for annotation. Thus, the best approach for future research may be to combine rule-based and ML methods.

6.2 Directions for Future Research

As discussed in 2, online book reviews are not necessarily representative of the unmediated reading experience, let alone of the spectrum of reading experiences that a book may evoke in different readers. Given the increasing amount of work that uses online book response in the study of reading, research that bridges these gaps seems particularly urgent, for ourselves as well as for the wider field of research in reading and reception.

Another ambitious next step in studying reading impact is the possibility of connecting the impact reported in online book reviews to specific features of individual books. This is the aim of the *Impact and Fiction* project (https://impactandfiction.huygens.knaw.nl/), where we will develop new metrics to computationally identify high-level features of literary texts such as mood, style and narrative structure, in order to examine the relationship between these book-intrinsic features and the impact of books expressed in online reviews. Additionally, we will differentiate between groups of readers to take into account that different groups of readers respond to book features in different ways (Van den Hoven et al. 2016). The research presented in this paper serves as a first step towards answering such questions.

7. Data Availability

Data can be found here: https://zenodo.org/record/5798598

8. Software Availability

Software can be found here: https://github.com/marijnkoolen/reading-impact-model/

9. Acknowledgements

10. Author Contributions

Marijn Koolen: Formal Analysis, Conceptualization, Software, Writing – original draft, Writing – review & editing

Julia Neugarten: Formal Analysis, Conceptualization, Writing – original draft, Writing – review & editing

Peter Boot: Formal Analysis, Conceptualization, Writing – original draft, Writing – review & editing

References

Alm, Cecilia Ovesdotter and Richard Sproat (2005). "Emotional sequencing and development in fairy tales". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, 668–674. 10.1007/11573548_86.

Basile, Valerio (2020). “It’s the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks”. In: *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*. Vol. 2776. CEUR-WS, 31–40. http://ceur-ws.org/Vol-2776/paper-4.pdf (visited on 10/24/2022).

Belfiore, Eleonora and Oliver Bennett (2007). “Determinants of impact: Towards a better understanding of encounters with the arts”. In: *Cultural trends* 16 (3), 225–275.

Boot, Peter and Marijn Koolen (2020). “Captivating, splendid or instructive? Assessing the impact of reading in online book reviews”. In: *Scientific Study of Literature* 10.1, 66–93. 10.1075/ssol.20003.boo.

Champagne, Ashley (2020). “What Is A Reader? The Radical Potentiality of Goodreads to Disrupt the Literary Canon”. In: *Digital Humanities 2020*. 10.17613/9rgd-t056.

Dadvar, Maral, Claudia Hauff, and Franciska De Jong (2011). “Scope of negation detection in sentiment analysis”. In: *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011)*. Citeseer, 16–20. 10.1016/j.dss.2016.05.009.

De Smedt, Tom and Walter Daelemans (2012). “’Vreselijk mooi!’(terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, 3568–3572. http://www.lrec-conf.org/proceedings/lrec2012/pdf/312_Paper.pdf (visited on 10/24/2022).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint*. 10.48550/arXiv.1810.04805.

Ehrmanntraut, Anton, Thora Hagen, Leonard Konle, and Fotis Jannidis (2021). “Type- and Token-based Word Embeddings in the Digital Humanities”. In: *CHR 2021, Proceedings of the Conference on Computational Humanities Research 2021*. CEUR-WS, 16–38. http://ceur-ws.org/Vol-2989/long_paper35.pdf (visited on 10/24/2022).

Gordon, Mitchell L, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein (2021). “The disagreement deconvolution: Bringing machine learning performance metrics in line with reality”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. 10.1145/3411764.3445423.

Green, Melanie C and Timothy C Brock (2000). “The role of transportation in the persuasiveness of public narratives.” In: *Journal of Personality and Social Psychology* 79 (5), 701–721.

Holur, Pavan, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R Tangherlini, and Vwani Roychowdhury (2021). “Modeling Social Readers: Novel Tools for Addressing Reception from Online Book Reviews”. In: *arXiv preprint*. 10.48550/arXiv.2105.01150.

Hutto, Clayton and Eric Gilbert (2014). “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. 8. https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (visited on 10/24/2022).

Kenyon-Dean, Kian, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths (2018). “Sentiment analysis: It’s complicated!” In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers)*, 1886–1895. 10.18653/v1/N18-1171.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel (2014). "The Sketch Engine: ten years on". In: *Lexicography*, 7–36.

Klenner, Manfred, Anne Göhring, Michael Amsler, Sarah Ebling, Don Tuggener, Manuela Hürlimann, and Martin Volk (2020). "Harmonization sometimes harms". In: *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. 0.5167/uzh-197961.

Knoop, Christine A, Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus (2016). "Mapping the aesthetic space of literature "from below"". In: *Poetics* 56, 35–49. 10.1016/j.poetic.2016.02.001.

Koolen, Marijn, Peter Boot, and Joris J. van Zundert (2020). "Online Book Reviews and the Computational Modelling of Reading Impact". In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, 149–169. http://ceur-ws.org/Vol-2723/long13.pdf (visited on 10/24/2022).

Koopman, Eva Maria Emy (2016). "Effects of 'Literariness' on Emotions and on Empathy and Reflection after Reading". In: *Psychology of Aesthetics, Creativity, and the Arts* 10 (1), 82–98.

Koopman, Eva Maria Emy and Frank Hakemulder (2015). "Effects of literature on empathy and self-reflection: A theoretical-empirical framework". In: *Journal of Literary Theory* 9 (1), 79–111. 10.1515/jlt-2015-0005.

Kuhn, Axel (2015). "2.3.3 Lesen in digitalen Netzwerken". In: *Lesen: Ein interdisziplinäres Handbuch*. Ed. by Ursula Rautenberg and Ute Schneider. De Gruyter, 427–444.

Kuijpers, Moniek M, Frank Hakemulder, Ed S Tan, and Miruna M Doicaru (2014). "Exploring absorbing reading experiences." In: *Scientific Study of Literature* 4 (1). 10.1075/ssol.4.1.05kui.

Lindell, Michael K and Christina J Brandt (1997). "Measuring interrater agreement for ratings of a single target". In: *Applied Psychological Measurement* 21 (3), 271–278. 10.1177/01466216970213006.

Miall, David S and Don Kuiken (2002). "A Feeling for Fiction: Becoming What We Behold". In: *Poetics* 30 (4), 221–241.

Miller, Christopher R (2015). *Surprise: The poetics of the unexpected from Milton to Austen*. Ithaca, NY: Cornell University Press. 10.7591/9780801455780.

Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov (2020). "Evaluation of sentiment analysis in finance: from lexicons to transformers". In: *IEEE Access* 8, 131662–131682.

Nell, Victor (1988). *Lost in a Book: The Psychology of Reading for Pleasure*. Yale University Press. 10.2307/j.ctt1ww3vk3.

Oortwijn, Yvette, Thijs Ossenkoppele, and Arianna Betti (2021). "Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 131–141. https://aclanthology.org/2021.humeval-1.15 (visited on 10/24/2022).

Poletti, Anna, Judith Seaboyer, Rosanne Kennedy, Tully Barnett, and Kate Douglas (2016). "The affects of not reading: Hating characters, being bored, feeling stupid". In: *Arts and Humanities in Higher Education* 15 (2), 231–247. 10.1177/1474022214556898.

Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C

Messerli, et al. (2019). *Digital humanities and digital social reading*. OSF Preprints. 10.31219/osf.io/mf4nj.

Ross, Catherine Sheldrick (1999). "Finding without Seeking: the Information Encounter in the Context of Reading for Pleasure". In: *Information Processing & Management* 35 (6), 783–799.

Sabine, Gordon and Patricia Sabine (1983). *Books That Made the Difference: What People Told Us*. ERIC.

Schmidt, Thomas, Manuel Burghardt, and Katrin Dennerlein (2018). "Sentiment annotation of historic german plays: An empirical study on annotation behavior". In: *Proceedings of the Workshop on Annotation in Digital Humanities 2018 (annDH 2018)*. RWTH Aachen. http://ceur-ws.org/Vol-2155/schmidt.pdf (visited on 10/24/2022).

Schmitt-Matzen, Cassie D (2020). "Adult Retrospectives on Unhealthy Adolescent Responses to Reading Fiction". PhD thesis. Tennessee Technological University.

Shannon, Donna M (1993). "Children's responses to humor in fiction". PhD thesis. The University of North Carolina at Chapel Hill.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170.pdf (visited on 10/24/2022).

Spiteri, Louise F and Jen Pecoskie (2016). "Affective taxomonies of the reading experience: Using user-generated reviews for readers' advisory". In: *Proceedings of the Association for Information Science and Technology* 53 (1), 1–9. 10.1002/pra2.2016.14505301032.

Sprugnoli, Rachele, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti (2016). "Towards sentiment analysis for historical texts". In: *Digital Scholarship in the Humanities* 31 (4), 762–772. 10.1093/llc/fqv027.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas (2010). "Sentiment strength detection in short informal text". In: *Journal of the American Society for Information Science and Technology* 61 (12), 2544–2558. 10.1002/asi.21416.

Thelwall, Mike and Kayvan Kousha (2017). "Goodreads: A social network site for book readers". In: *Journal of the Association for Information Science and Technology* 68 (4), 972–983. 10.1002/asi.23733.

Tobin, Vera (2018). *Elements of surprise: Our mental limits and the satisfactions of plot*. Cambridge, MA and London, England: Harvard University Press. 10.4159/9780674919570.

Usherwood, Bob and Jackie Toyne (2002). "The value and impact of reading imaginative literature". In: *Journal of Librarianship and Information science* 34 (1), 33–41.

Van den Hoven, Emiel, Franziska Hartung, Michael Burke, and Roel M Willems (2016). "Individual differences in sensitivity to style during literary reading: Insights from eye-tracking". In: *Collabra* 2 (1). 10.1525/collabra.39.

Wan, Mengting and Julian J. McAuley (2018). "Item recommendation on monotonic behavior chains". In: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys*. ACM, 86–94. 10.1145/3240323.3240369.

Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley (2019). "Fine-Grained Spoiler Detection from Large-Scale Review Corpora". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Vol 1: Long*

Papers. Association for Computational Linguistics, 2605–2610. 10.18653/v1/p19-1248.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo (2019). "Evaluating word embedding models: Methods and experimental results". In: *APSIPA transactions on signal and information processing* 8. 10.1017/ATSIP.2019.12.

Wang, Yile, Leyang Cui, and Yue Zhang (2019). "How Can BERT Help Lexical Semantics Tasks?" In: *arXiv preprint*. 10.48550/arXiv.1911.02929.

Wu, Fangzhao and Yongfeng Huang (2016). "Sentiment domain adaptation with multiple sources". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Vol. 1: Long Papers*, 301–310. 10.18653/v1/P16-1029.

Yao, Fang and Yan Wang (2020). "Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach". In: *Computers, Environment and Urban Systems* 83. 10.1016/j.compenvurbsys.2020.101522.