Detection of Reading Absorption in User-Generated Book Reviews:Resources Creation and Evaluation

Piroska Lendvai, Sándor Darányi, Christian Geng, Moniek Kuijper, Oier Lopez de Lacalle, Jean-Christophe Mensonides, Simone Rebora, Uwe D. Reichel

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4835–4841
Marseille, 11–16 May 2020

Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation

Piroska Lendvai[1], Sándor Darányi[2], Christian Geng[3], Moniek Kuijpers[1], Oier Lopez de Lacalle[4], Jean-Christophe Mensonides[5], Simone Rebora[1], Uwe Reichel[6]
[1]DH Lab, University of Basel, Switzerland [2]University of Borås, Sweden
[3]Carstens Medizinelektronik GmbH, Germany [4]University of the Basque Country, Spain
[5]LGI2P, IMT Mines Ales, France [6]Research Institute for Linguistics, Hungary
Corresponding author: piroska.r@gmail.com

Abstract

To detect how and when readers are experiencing engagement with a literary work, we bring together empirical literary studies and language technology via focusing on the affective state of absorption. The goal of our resource development is to enable the detection of different levels of reading absorption in millions of user-generated reviews hosted on social reading platforms. We present a corpus of social book reviews in English that we annotated with reading absorption categories. Based on these data, we performed supervised, sentence level, binary classification of the explicit presence vs. absence of the mental state of absorption. We compared the performances of classical machine learners where features comprised sentence representations obtained from a pretrained embedding model (Universal Sentence Encoder) vs. neural classifiers in which sentence embedding vector representations are adapted or fine-tuned while training for the absorption recognition task. We discuss the challenges in creating the labeled data as well as the possibilities for releasing a benchmark corpus.

Keywords: social reading corpus, sentence embeddings, reading absorption detection, affective computing

1. Introduction

Our study aims to contribute to text-based affect recognition research by focusing on the experience-related state of absorption during reading literary fiction. Our goal was to process user-generated book reviews from an online social reading platform, and detect passages that textually express reading absorption, e.g. in terms of phrases such as *'I was completely hooked'*. The reviews belong to the genre of non-elicited self-narratives that emerge in online social reading communities; cf. (Cordón-García et al., 2013; Rebora et al., 2019). They are subjective, opinionated, unstructured self-reports of variable length that typically include references to one's individual reading experience. As opposed to product reviews, the texts often do not merely contain mentions of evaluative sentiment toward (components of) a book, but also express complementary, affective aspects such as reviewers' cognitive engagement during their individual reading experience.

Reading absorption (aka narrative absorption) has traditionally been investigated in the humanities and social sciences (Hakemulder et al., 2017), e.g. by empirical literary and aesthetics studies. Importantly, absorption has been found to be composed of multiple facets, such as transportation to the fictional world ("*I feel like I've just returned from a long vacation in Martin's fantasy kingdom*"), focused attention ("*a wonderful one that really draws you in*"), altered sense of time during reading (*"It went by in a blink"*), emotional engagement (*"I instantly connected with Annabel"*), and others.

1.1. Previous work

There have been empirical studies on absorption during reading that have focused on the textual determinants of such experiences (Green and Brock, 2000), the individual differences that can predict the occurrence of such experiences in specific types of readers (Kuijpers et al., 2019), and the outcomes of absorbed reading in terms of persuasive (Green and Brock, 2000) or aesthetic effects (Kuiken and Douglas, 2017). These experiments were however conducted in controlled laboratory settings, while absorption is an experience that is hard to simulate in a lab. Reader reviews on social sites on the other hand contain information on absorbing reading experiences that emerged naturally.

Computational detection of reader absorption has so far been largely unaddressed by the Natural Language Processing (NLP) community. Unlike detecting emotion, opinion, polarity, sentiment, stance, subjectivity, (Liu, 2012; Zhang et al., 2015; Mohammad et al., 2016; Balahur et al., 2018), the absorption detection task involves interpreting a complementary, affective state of reader response. Its distinctive features are subtle and not yet fully explored, which seems to challenge even trained human annotators.

We presented our previous approaches to automatically detect absorption in the story world, using two NLP methods: textual entailment detection and text reuse detection (Rebora et al., 2018a; Rebora et al., 2018b). The task we constructed was to detect semantic similarity between the 18 statements in the *Story World Absorption Scale (SWAS)* questionnaire and cca. 3,500 sentences in Goodreads reviews, which we manually labeled. Table 1 shows part of the 18 statements of the SWAS), an instrument developed by (Kuijpers et al., 2014) for the purpose of investigating reading experience in the field of empirical literary studies.

Textual entailment detection was conducted using the Excitement Open Platform (Magnini et al., 2014), i.e. a maximum entropy classifier utilizing WordNet, VerbOcean, bag-of-dependencies scoring using TreeTagger, and tree skeleton scoring; for details cf. (Wang and Neumann, 2007). We

Component	Statement
...	...
Transportation	When I was reading the story it sometimes seemed as if I were in the story world too
	When reading the story there were moments in which I felt that the story world overlapped with my own world
	The world of the story sometimes felt closer to me than the world around me
	When I was finished with reading the story it felt like I had taken a trip to the world of the story
	Because all of my attention went into the story, I sometimes felt as if I could not exist separate from the story
Emotional Engagement	When I read the story I could imagine what it must be like to be in the shoes of the main character
	I felt sympathy for the (main) character(s)
	I felt connected to the (main) character(s) in the story
	I felt how the (main) character(s) was/were feeling
	I felt for what happened in the story
...	...

Table 1: Part of the SWAS absorption questionnaire created in empirical literary studies (Kuijpers et al., 2014).

used both the pretrained classifier and also retrained the tool on a small set of 480 balanced entailment pairs created from our social reading data. Text reuse detection was conducted on the same pairs with TRACER (Franzini et al., 2018) using token-level preprocessing with synonyms and hyponyms from WordNet, a 16-word moving window and other features. Neither of these two approaches yielded an F-score above 0.10 on the target class (i.e., 'entailment' or 'similarity' vs. 'non-entailment' or 'non-similarity').
We also targeted the text-based identification of generic reading absorption in pilot experiments using baseline supervised text classification approaches with logistic regression and random forests as reported in (Lendvai et al., 2019). Our pilot experiments were based on a self-created corpus of 200 reviews, annotated by 5 trained raters, based on which we obtained encouraging results: an F-score of .42 on detecting the target class *Absorption* on the sentence level with the base classifiers and simple text-based token count representation. For these experiments, we generated our own sentence embedding model based on 2.5 million user reviews from the Goodreads platform by retraining the *sent2vec* tool's model (Pagliardini et al., 2018), which did not provide a better-performing content representation in the pilot experiments. We are not aware of further computational work on absorption detection.

1.2. Contributions of this study

The SWAS remained at the core of a larger absorption inventory developed for the current project and used for annotating specific and related instances of absorption in reader reviews. In contrast to our previous investigations, in the current study we were interested in identifying the broad affective state of *reading absorption* next to absorption in the story world. Therefore, we incorporated in the labeling and prediction task additional absorption concepts such as participatory responses to the fictional characters (*"I want to be Molly when I grow up"*), lingering story feelings ("*leaves you with goosebumps after having read the last page"*), etc. We extended our experimental investigation by constructing a larger corpus and performing absorption detection using state-of-the-art classification methods. In particular:

1. We present a corpus of 380 social reading reviews in English that are manually annotated with reading absorption categories

2. We explore three sentence representation models for absorption detection

   (a) We use sentence embeddings from the Universal Sentence Encoder (Cer et al., 2018) and feed them to classical machine learners for the end task.

   (b) We train fastText (Joulin et al., 2016) on our corpus to learn sentence embeddings while training on the end task

   (c) We perform transfer learning by employing pretrained deep bidirectional sentence embeddings of BERT (Devlin et al., 2018), which we fine-tune on our supervised end task using our social reading corpus data.

Via the different learning approaches and corresponding evaluative experiments, we aim to gain insight into affective phenomena in user-generated opinionated texts that are complementary to conventional emotional categories, which can benefit computational linguistics, literature and social sciences studies, as well as industrial content analysis.

2. Corpus construction

Our current corpus is constructed using 380 English review texts for 224 books, which we collected from a social reading platform. These reviews pertained to books from different literary genres (romance, fantasy, science fiction, thriller) that we pre-selected based on high star-ratings on the platform and the presence of trigger words. The majority of the books (180) had a single review in the corpus, 33 books had 5 reviews each, the rest of the books had a small number of reviews (<7).

2.1. Manual annotations

We have been extensively instructing 5 in-house annotators with background in computational linguistics and/or literary studies for labeling reading absorption. The set of labels was weakly structured in terms of broad absorption concepts such as Attention, Transportation, Emotional Engagement, Mental Imagery, Disconnection from reality, etc. (cf. Table 1), which each held narrow absorption concepts taken from the inventories listed and discussed by (Kuijpers et al., 2014) and (Bálint et al., 2016). The label set totaled about 40 distinct concepts (cf. Table 2); we note that the labeling scheme is still subject to revision.
The annotators could assign the labels to text segments of any length within a review. The criteria for assigning preferably one label (but possibly more) was driven by, but not restricted to, the semantic similarity between some text segment and the statements or concepts in the inventories (cf.

Sent	Text	**Fine-grained label** (nr Annotations)	**Binary task label**
1	The first time I tried to read Neuromancer, I stopped around page 25.	Unwillingness to stop reading – Lack (1)	Abs_min
2	I was about 15 years old and I'd heard it was a classic, a must-read from 1984.	-	Nonabs
3	So I picked it up and I plowed through the first chapter, scratching my head the whole time.	Effortless engagement – Lack (1)	Abs_min
...	...	...	...
7	This time, William Gibson's dystopic rabbit hole swallowed me whole.	Attention (2), Transportation (1), Effortless engagement (1)	Abs_maj
...	...	...	...
25	No, you're thrust right into Case's shoes as he swills rice beer in Japan and pops amphetamines and tries to con the underworld in killing him when his back is turned because he thinks he'll never work again.	Emotional engagement (1)	Abs_min
26	You have to piece together the rest on your own.	-	Nonabs
27	Challenging?	Effortless engagement – Lack (2)	Abs_min
28	You bet.	Effortless engagement – Lack (2)	Abs_min
29	But it's electrifying once you get it.	-	Nonabs
30	I've worked by paperback copy until the spine and cover have split, until the pages have faded like old newsprint.	-	Nonabs
31	Echoes of its diction sound in my own writing.	-	Nonabs
32	Thoughts of Chiba City or BAMA pop into my head when I walk through the mall and hear a melange of voices speaking in Spanish and English and Creole and German.	Lingering story feelings (2)	Abs_min
33	Neuromancer is in me like a tea bag, flavoring my life, and I ca	Lingering story feelings (3)	Abs_maj
34	n't imagine what it would be like if I hadn't pressed on into page 26.	Lingering story feelings (2)	Abs_min

Table 2: Corpus excerpt of automatically segmented sentences, with manual absorption categories annotated (column 3) and the binarized target classes (column 4). *Abs_min*: reading absorption or its lack was explicitly expressed by at least 1 assigned absorption label; *Abs_maj*: by at least 3 absorption labels assigned; *Nonabs*: no absorption or its lack was labeled. Note that erroneous sentence segmentations (e.g. between sentences 33 and 34) resulted from automatic sentence splitting.

Table 1). The boundaries of the relevant text snippet were to be freely established by each annotator, as the raters were presented full review texts using the Brat annotation platform (Stenetorp et al., 2012). The annotators could also mark it up when users explicitly mentioned or signaled the lack of absorption (e.g. *"I struggled to get through a lot of the pages*" or *"None of the characters really mattered to me"*), to make them distinct from expressions that describe the presence of absorption.

In Table 2, we illustrate a review excerpt where the central column for each sentence shows the manually assigned absorption labels and the number of annotators assigning that label. Note that sentences 29-31 were not labeled as showing absorption, since these sentences carry evaluation or sentiment related to the generic reading experience or its impact on the reviewer, rather than reporting about having been (or not) in a specific state of absorption as defined by us (cf. the broad absorption concepts in our labeling scheme). These specific states are often expressed in the reviews in terms of linguistically distinct phrases, and our project goal was to capture such direct expressions of absorption. The annotators were therefore explicitly instructed not to mark up passages that express only speculatively inferrable reading absorption or generic sentiment. We give an overview of the annotation process in (Rebora et al., 2020).

2.2. Post-processing

After the completion of manual annotations, we performed the following steps.

Text normalization To reduce the noisiness of the review texts, we normalized character encoding using *unidecode*[1] and a series of simple character mapping rules. We also replaced emoticons with their descriptions[2] and masked full URLs with a placeholder.

Sentence-level labeling normalization To construct the dataset for classification experiments, we mapped all annotated text spans to the sentence level.

Sentence boundary segmentation was obtained after the annotations were completed, using the *spaCy* package[3]. In the corpus, the mean review length was 23.9 sentences, and the mean sentence length was 15.5 tokens.

Labeling adjudication and inter-annotator agreement On the current corpus we obtained a 0.40 sentence-level Krippendorff's Alpha inter-annotator agreement score. Adjudication of the labels obtained from the annotators is currently ongoing. A field expert created gold-standard labels so far for 1,475 sentences, originating from more than 60 reviews, which yielded a 0.59 Cohen's Kappa mean score.

2.3. Binary class partitioning

Since at this point in our project, labeled evidence for the individual absorption categories turned out to be small, for the current study we aggregated all absorption types into a generic class *Absorption* as opposed to *Nonabsorption*. The *Absorption* class was constructed according to two separate threshold values:

1. the *Abs_maj* class was assigned if at least three annotations of any absorption kind were marked up in some part of the sentence i.e., if the labels assigned by all annotators summed to at least 3.

2. the *Abs_min* class was assigned if at least one annotator assigned a label to some part of the given sentence, i.e., if all labels that were assigned by all the annotators summed to at least 1.

These choices were motivated by the moderate inter-annotator agreement about label identity of the fine-grained absorption categories, and the moderate amount of currently available data. Table 2 shows both the manually assigned

[1] https://pypi.org/project/Unidecode/

[2] https://en.wikipedia.org/wiki/List_of_emoticons

[3] https://spacy.io

Model	Majority vote task						Low threshold task					
	Abs			Nonabs			Abs			Nonabs		
	P	R	F	P	R	F	P	R	F	P	R	F
Baseline: majority class	0	0	0	.95	1.0	.98	0	0	0	.85	1.0	.92
Random forest, USE vec	.56	.04	.07	.95	.99	.98	.65	.12	.21	.87	.99	.92
SVM rbf, USE vec	.12	.76	.20	.98	.72	.83	.27	.63	.37	.91	.7	.79
SVM linear, USE vec	.17	.74	.27	.98	.82	.89	.31	.69	.43	.93	.74	.83
Logistic regression, USE vec	.17	.74	.28	.98	.82	.90	.32	.68	.43	.93	.74	.83
Gradient boosting, USE vec	.32	.24	.27	.96	.98	.97	.45	.45	.45	.90	.90	.90
fastText, scratch	.33	.34	.33	.98	.97	.97	.29	.51	.37	.90	.78	.84
fastText, finetuned	.40	.30	.34	.97	.98	.97	.39	.37	.38	.89	.90	.89
BERT, frozen	.07	.75	.14	.98	.55	.70	.21	.70	.32	.91	.53	.67
BERT, finetuned	.58	.34	**.43**	.97	.99	.98	.57	.51	**.54**	.92	.93	.92

Table 3: Classification results with 5-fold CV and oversampling for the two tasks. Evaluation is in terms of Precision (P), Recall (R) and F1-score (F).

absorption labels as well as the binarized labels that were used in the classification experiments.

3. Detecting reading absorption: Experimental setup

Our two classification tasks consisted of binary decisions:

1. Majority vote task: separating the *Abs_maj* sentences (422) from the merged *Nonabs* + *Abs_min* sentences (8,653)

2. Low threshold task: separating the merged *Abs_min* + *Abs_maj* sentences (1,339 instances) from *Nonabs* sentences (7,736).

We tested classical feature-based algorithms, as well as neural algorithms on these tasks. Only generic values were passed for hyperparameters, i.e. no fine tuning was performed for any of them. The information the classifiers drew on are only the sentence representations.

3.1. Data partitioning

The dataset is heavily imbalanced in both tasks. Without oversampling, performance on the Abs class is often 0.We used the random oversampling method from the *imbalanced-learn* package[4] during training to remedy this. We conducted all experiments via 5-fold cross-validation.

3.2. Sentence representation

Neural models can extract high-quality sentence representations, pretrained on large amounts of unlabeled texts. In the conventional classifiers, we directly imported the representations as generated by the pretrained Universal Sentence Encoder model (USE); these are 512-dimensional embedding vectors.

3.3. Classical ML models

We used four conventional machine learners from the scikit-learn library(Pedregosa et al., 2011) to which we fed sentence representations obtained from the Universal Sentence Encoder (USE) as implemented in TensorFlow. Each learner was run without parameter optimization. For the Random Forest and the Gradient boosting classifier the number of estimators was set to 500. For all other hyperparameters *sklearn* default values were used.

The results for these models are presented in the upper half of Table 3. The scores show that in the Majority vote task where we have very little and imbalanced data for our class of interest (422 positive instances altogether, all of them unique except for one case), F scores for all models stay below .30 points, and the gradient boosting, logistic regression and linear SVM models perform similarly.

In the Low threshold task, where we have 3 times as many positive instances (nearly all unique), the performance of these three algorithms is best again, in the .43-.45 F-score interval.

3.4. Neural models

We used two neural learning tools: fastText (Joulin et al., 2016) and BERT (Devlin et al., 2018), each with their native encoders and their standard settings. The employed neural classifiers created representation models from scratch (fastText) or implemented different approaches for the supervised end-task after embedding initialization, the latter by either freezing pretrained layers and only training a final layer (BERT) or allowing for updating the entire model (fastText, BERT).

The results are presented in the lower half of Table 3.

3.4.1. fastText

We trained two fastText models. The 'fastText scratch' model did not involve pretraining-based initialization, but was trained fully on our data, while 'fastText finetuned' used a pretrained encoder model updated while using our data.

3.4.2. BERT

BERT bidirectionally learns contextual representations for words. In 'BERT frozen', representations are generated based on a pretrained encoder model and not get updated on our data, whereas the 'BERT fine-tuned' case involves a pretrained encoder that is updated based on our data, analogous to the 'fastText fine-tuned' case. We used the pytorch interface for BERT [5] and the pretrained bert-base-uncased model.

[4] https://github.com/scikit-learn-contrib/imbalanced-learn

[5] https://github.com/huggingface/transformers

3.5. Results and discussion

The experimental outcomes suggest that all models still suffer from data imbalance, even though oversampling is applied. The models are trained on textual data only, in which absorption is not straightforward to detect even for trained humans, which is expressed in moderate inter-annotator agreement. Overall, the results are promising, especially for BERT-finetuned, which having an imbalanced dataset is able to obtain an F-score of 43 points for the most infrequent class in the Majority vote labeling task, and 54 points in the Low threshold labeling task. Regarding the vector dimensionality of each representation model, we assume that each model has its own optimum size. Regarding model comparison, we observe the following:

- The fine-tuned version of BERT outperforms the rest of the models. This might be since, in accordance with previous studies, this model is able to provide highly contextualized representations, obtaining very good results on NLP end tasks. The BERT frozen model is not able to show as strong a generalization capability however, with a surprisingly low recall in the Majority vote task, a dramatic difference between its performance in the finetuned setup. Transferring features without fine-tuning seems not to be helpful.

- The statically encoded vector models from fastText are not able to provide as good a representation as BERT. On the Low threshold task the fastText performances stay even below the conventional classifiers' scores.

- The USE encodings are able to provide a stronger semantic representation of reading absorption than the fastText pretrained encoding model that we used. Still, the best score based on USE embeddings, from the gradient boost model, lags behind BERT with nearly a 10-point difference (.45 F, Low threshold task).

The difficulty of textual assessment of absorption can be illustrated by the sentence similarity matrices based on USE representations. In Figure 1 we present the matrix for the last 20 items from the *Abs_maj*-labeled sentences, whereas in Figure 2 the matrix for the full set of 466 sentences in the Majority vote task, and in Figure 3 the matrix for the full set of 1339 sentences in the Low Threshold task. The heatmaps for the positive class do not uniformly reflect high semantic homogeneity across the sentences. Note that the label propagation from arbitrary text segments to sentences introduces noise, and while the similarity matrix and the classification experiments are set up on the (observably often erroneous) sentence level, most of our annotated chunks span less than an entire sentence.

In line with our task setup, for the Low threshold condition the sentences labelled as Absorption show a lower degree of similarity. For the Majority vote condition, sentence similarity within the smaller set increases, but so does the data imbalance problem. Thus, performance appears to be limited for both tasks, either by the lack of high quality textual cues or by a skewed class distribution.

4. Summary and outlook

Goals and workflow We described the creation of language technology resources that enable the detection of reading absorption in textual data from online social platforms. To create the resources, we annotated 380 reviews in terms of a large set of absorption categories taken from or inspired by empirical literary studies. The current corpus features moderate inter-annotator agreement, demonstrating that the manual labeling task requires extensive effort comparable to previous findings (Kim and Klinger, 2018), and providing feedback for the adaptation of the annotation scheme. These efforts are currently ongoing.

To remedy the issue of still relatively small size as well as the uncertainty of the currently prepared labeled data, we recast the absorption detection task as binary classification. We tested different label construction thresholds and different types of classifiers for the end task. The features that the classifiers employed were sentence or word embedding vectors from different language models.

Experimental findings and ongoing work The experimental outcomes suggest that USE encodings are able to provide a robust semantic representation of reading absorption on par for this task with the statically encoded vector models from fastText, while the fine-tuned version of BERT outperforms the rest of the models. Data imbalance could partly be remedied using oversampling, while – based on our two tasks that used different inter-annotator agreement thresholds – we assumed that even lower quality annotations can benefit BERT when available in large amounts.

Social reading self-narratives incorporate multiple complementary affective phenomena, such as subjectivity and sentiment, as well as multiple user intents besides reporting reading absorption, e.g. evaluation, recommendation, feedback, socializing. Therefore, our ongoing work addresses the incorporation and evaluation of these phenomena in a complex absorption detection model.

Language resources output The possibilities for releasing our resources depend on tackling several challenges. We attempted to clarify the legal aspects regarding the collection, processing and dissemination of the data from the Goodreads platform. The information currently available to us does not allow reuse of these texts, which poses a risk to freely sharing our 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. These exceptions have been already included in national laws such as the *Urheberrechts-Wissensgesellschafts-Gesetz* in Germany[6]. However, this law states that one should either destroy or store the data in a protected repository after analysis. We are evaluating the possibility of making a consolidated benchmark corpus accessible under a specific license, after having complied with all legal and ethical requirements.

5. Acknowledgments

Lendvai, Rebora and Kuijpers were supported by the Swiss National Science Foundation within the *Mining Goodreads*

[6] https://www.clarin.eu/content/clic-copyright-exceptions-germany

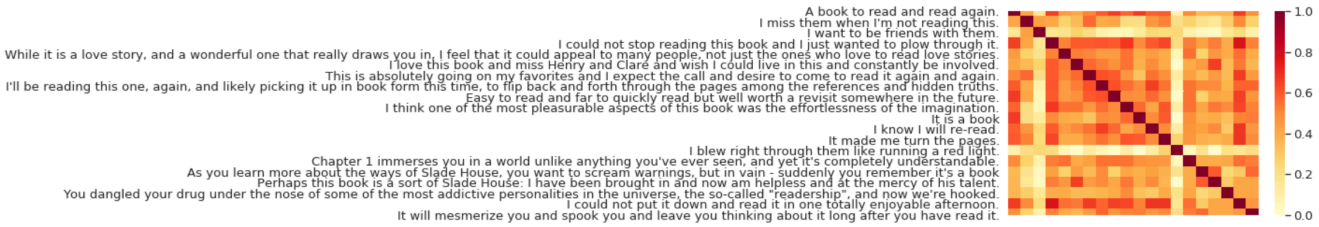

Figure 1: Cosine similarity matrix across the last 20 sentences in the corpus that were labeled as *abs_maj*, represented as 512-dimensional sentence embedding vectors from the pretrained Universal Sentence Encoder model. (Sentences are cropped due to space constraints.)

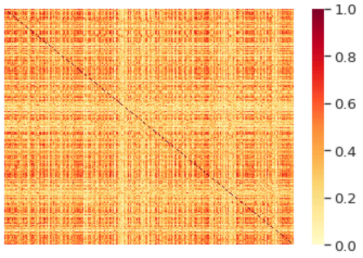

Figure 2: Cosine similarity matrix based on USE embeddings across all 422 sentences in the corpus that were labeled as *abs_maj*.

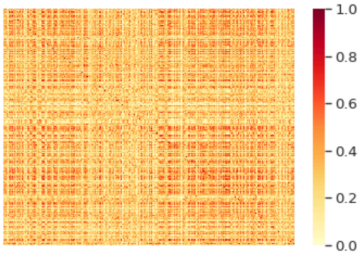

Figure 3: Cosine similarity matrix based on USE embeddings across all 1,339 sentences in the corpus that were labeled as *abs_min*.

project (Grant nr. 10DL15_183194). We would like to thank Ute Winchenbach (Technische Universität Darmstadt, Germany) for coding the BERT wrapper and for suggestions for improving the training workflow.

6. Bibliographical References

Balahur, A., Mohammad, S., Hoste, V., and Klinger, R. (2018). Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

Bálint, K., Hakemulder, F., Kuijpers, M., Doicaru, M., and Tan, E. S. (2016). Reconceptualizing foregrounding. *Scientific Study of Literature*, 6(2):176–207.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Cordón-García, J.-A., Alonso-Arévalo, J., Gómez-Díaz, R., and Linder, D. (2013). Social reading: platforms, applications, clouds and tags. Elsevier.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Franzini, G., Franzini, E., Bulert, K., Büchler, M., and Moritz, M. (2018). Tracer: A user manual.

Green, M. C. and Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701.

Hakemulder, F., Kuijpers, M. M., Tan, E. S., Bálint, K., and Doicaru, M. M. (2017). Narrative absorption, volume 27. John Benjamins Publishing Company.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv: 1607.01759*.

Kim, E. and Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1345–1359. Association for Computational Linguistics.

Kuijpers, M. M., Hakemulder, F., Tan, E. S., and Doicaru, M. M. (2014). Exploring absorbing reading experiences. *Scientific Study of Literature*, 4(1).

Kuijpers, M., Douglas, S., and Kuiken, D. (2019). Personality traits and reading habits that predict absorbed narrative fiction reading. *Psychology of Aesthetics, Creativity, and the Arts*, 13(1):74.

Kuiken, D. and Douglas, S. (2017). Forms of absorption that facilitate the aesthetic and explanatory effects of literary reading. *Narrative Absorption*, 27:219–252.

Lendvai, P., Rebora, S., and Kuijpers, M. (2019). Identification of reading absorption in user-generated book reviews. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019).

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Padó, S., Stern, A., and Levy, O. (2014). The Excitement Open Platform for Textual Inferences. In Proceedings of ACL Demo Session.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In Proceedings of the International

Workshop on Semantic Evaluation, SemEval '16, San Diego, California.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rebora, S., Kuijpers, M., and Lendvai, P. (2018a). Mining Goodreads: A text similarity based method to measure reader absorption. In Proceedings of the 3rd Swiss Text Analytics Conference (SwissText).

Rebora, S., Lendvai, P., and Kuijpers, M. (2018b). Reader experience labeling automatized: Text similarity classification of user-generated book reviews. In Proceedings of the European Association for Digital Humanities Conference (EADH).

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M., Lauer, G., Lendvai, P., and Messerli, T. C. (2019). Digital humanities and digital social reading. *OSF Preprints*.

Rebora, S., Lendvai, P., and Kuijpers, M. (2020). Annotating reader absorption. In: Proc. of Digital Humanities Conference (DH-2020).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (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, pages 102–107. Association for Computational Linguistics.

Wang, R. and Neumann, G. (2007). Recognizing textual entailment using a subsequence kernel method. In AAAI, volume 7, pages 937–945.

Zhang, M., Zhang, Y., and Vo, D. T. (2015). Neural Networks for Open Domain Targeted Sentiment. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 612–621. Association for Computational Linguistics.