To Reduce Bias, You Must Identify It First! Towards Automated Gender Bias Detection

Short Paper

Lorena Göritz
German Research Center for Artificial Intelligence, Parkstr. 40, 49080 Osnabrück, Germany,
lorena.goeritz@dfki.de

Daniel Stattkus
German Research Center for Artificial Intelligence, Parkstr. 40, 49080 Osnabrück, Germany,
daniel.stattkus@dfki.de

Jan Heinrich Beinke
German Research Center for Artificial Intelligence, Parkstr. 40, 49080 Osnabrück, Germany,
jan.beinke@dfki.de

Oliver Thomas
German Research Center for Artificial Intelligence,
Osnabrück University, Parkstr. 40, 49080 Osnabrück, Germany,
oliver.thomas@dfki.de

Abstract

Stereotypical gender representation in textbooks influences the personal and professional development of children. For example, if women do not pursue a STEM career because of gender stereotypes, this is not only an individual problem but also negative for society in general. It is hence crucial that textbooks do not convey gender stereotypes but are gender-balanced. Currently, textbook analysis is predominantly conducted manually, if at all. However, this is time-consuming and consequently cost-intensive. Therefore, as part of a design science research project, we developed a gender language analyzer. Our initial prototype is already capable of automatically analyzing textbooks and recommending suggestions regarding gender-balancing. We will further improve our prototype in the next design science research cycle (e.g., by integrating self-learning techniques). With this tool, publishers will be able to automatically analyze textbooks to reduce gender bias. Moreover, we provide the scientific community with design knowledge regarding automated identification of gender bias.

Keywords: gender bias, inclusion, natural language processing, design science research

Introduction

Children learn at an early age what their parents and teachers expect of them and what they think they can accomplish according to their gender. This is already apparent in the toys they are given to play with. While girls usually take care of their doll as if it were their baby, boys often play with superhero toys and footballs (Blakemore and Centers 2005). As part of this gender socialization, they develop strong hidden associations between objects and gender, which results in gender stereotypes (Miller 2010). Gender stereotypes can lead to negative consequences: girls and boys do what is expected of them, not what they truly excel at or enjoy. For example, a child is more likely to develop an interest in science if a positive correlation exists between the stereotypical scientist and the child's self-image (Hannover and Kessels 2004). This dynamic results in negative effects, both on a personal level, in the sense that people may be unhappy and in unrealized economic potential. Between the ages of 6 and 10, stereotype consciousness — the ability to infer a person's stereotype — increases significantly (McKown and Weinstein 2003). During this period, children spend a

lot of time at school, and the influence of school on the formation of gender stereotypes is consequently strong. Therefore, it is crucial to ensure that the learning environment in schools does not reinforce these stereotypes by using gender-biased teaching and assessment methods and educational resources (Kerkhoven et al. 2016). Previous research has already addressed this issue by manually analyzing textbooks for gender bias (Elgar 2004; Lee and Collins 2009; Moser and Hannover 2014). However, manual analysis is personnel- and time-intensive and hence results in high costs. To identify the reinforcement of gender stereotypes quickly and efficiently, replacing the previous manual analysis with an automated tool is inevitable. Such a tool would enable the rapid analysis of entire textbooks and not just individual paragraphs regarding gender bias. Furthermore, this tool could provide textbook authors with tips and guidance on gender-balancing. On the basis of these considerations, we derive our research question: *How can textbooks be automatically analyzed to detect gender bias?*

As part of a design science research (DSR) approach, we develop an initial prototype *Gender Language Analyzer (GLA)*. Our prototype is capable of automatically analyzing texts regarding gender bias and will be refined and improved in future development cycles, for example by integrating self-learning techniques. The first evaluation cycle provided evidence that the GLA already evaluates texts similarly to humans regarding gender bias. For companies (e.g., publishers), our prototype shares valuable insights into how tools can support the gender-balancing of textbooks. From a scientific point of view, our identified requirements, meta-requirements, and design objectives can provide starting points for future research. Our paper is structured as follows: we first explain the background of our research topic. Thereafter, we describe the research approach and derive requirements and design objectives for our GLA. We then describe the development of an initial GLA prototype and its evaluation. Finally, we discuss future changes and improvements.

Background

To identify gender bias in textbooks, the most frequently studied aspect is the prevalence of male and female characters in text and visuals (Elgar, 2004; Kerkhoven et al., 2016; Lee & Collins, 2009; Moser & Hannover, 2014). The design of course content in a stereotypically male or female context has also been examined (Kerger et al., 2011; Parkin & Mackenzie, 2017). A computer science course may use stereotypical male examples such as dragons and castles to vividly explain their content or examples from a stereotypical girls' world. Both options would exclude the opposite gender (Kerger et al. 2011). Therefore, depending on the purpose, it is essential to ensure that course materials are designed to be gender-balanced. Some authors have even made their methodological procedures for identifying gender bias in textbooks available as tools in the form of checklists and codebooks so that teachers, book editors, or researchers can use them to analyze their books and make them more gender-balanced (Parkin and Mackenzie 2017; De Waard and Zolfo 2009). However, these manual analyses are time-consuming, personnel-intensive, and therefore expensive.

In other domains, automated text analysis has already been investigated, for example the perception of job postings among men and women (Gaucher et al. 2011). Here, mainly two approaches have emerged: the most frequently used approach is the creation of a lexicon, and machine learning is proposed as an alternative. Using the lexical approach, words are mapped to specific subgroups such as male and female. We can further distinguish between lexica containing only words and lexica combining words into topics. For example, the male subgroup contains topics referring to violence, arrogance, sex, and strength (Fast et al. 2016). In the literature, the prevalent major approaches for developing lexica are the Bem Sex-Role Inventory (Bem 1974), which is one of the first approaches to create a lexicon for gender bias, and Cryan et al.'s (2020) approach, which involves asking men and women whether they would classify a word as male or female. Furthermore, Cryan et al. (2020) generated scores for words using supervised learning. The approach of developing a lexicon that combines words into topics is more frequently used than relying on individual words. It is used, for example, in the analysis of job postings (Gaucher et al. 2011), letters of recommendation (Madera et al. 2009), and the language of movie characters (Ramakrishna et al. 2017). The classification into topics is performed based on prior research (Fast et al. 2016; Gaucher et al. 2011; Wagner et al. 2015) and on existing tools such as LIWIC (Madera et al. 2009; Ramakrishna et al. 2017) and Empath (Sun et al. 2022). Cryan et al. (2020) highlight that lexica have a disadvantage because they lose relevancy over time. Therefore, the authors propose an end-to-end deep learning approach as an alternative.

Design Science Research Approach

Our research approach is based on the DSR framework of Kuechler and Vaishnavi (2008) and is illustrated in Figure 1, which is adapted from Diederich et al. (2020). To date, we have conducted the first design cycle. We interviewed 12 people working on the gender bias topic; four worked as professors or researchers in academia, and eight worked in industry as consultants, gender policy officers, or corporate communication officers. The interviews lasted between 37 and 70 minutes, and we coded the requirements inductively from the interviews (Mayring 2010). Furthermore, we reviewed the literature on the topic to refine and expand the requirements and formulate initial design objectives. After the interviews and literature review, we gathered a list of 12 requirements, seven meta-requirements, and four design objectives.

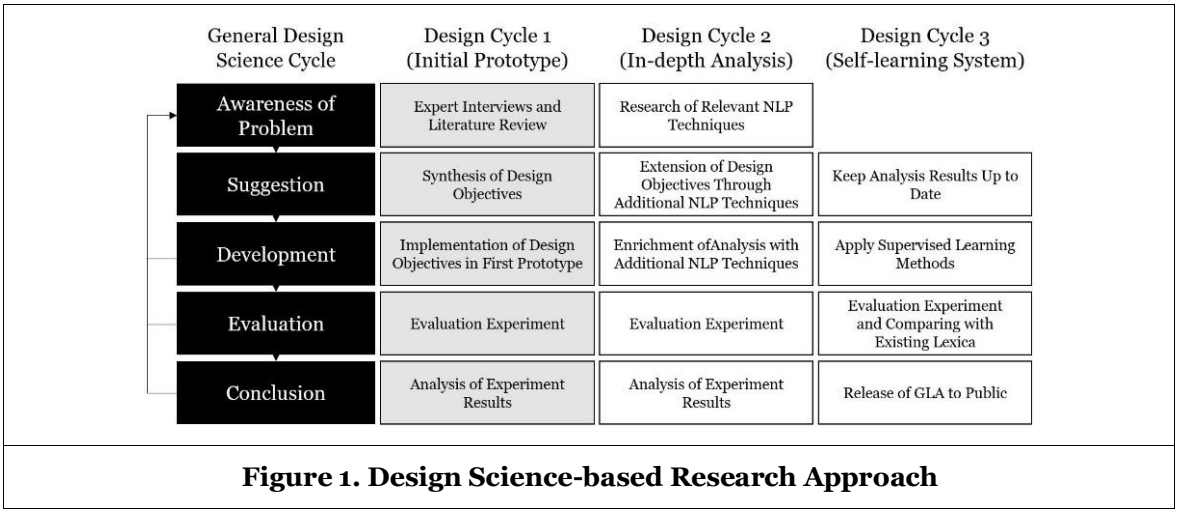

Figure 1. Design Science-based Research Approach

The design objectives were implemented in an early prototype including a lexicon, a gender-detection function for character names, and a dashboard that users could interact with. We used the CRISP-DM reference model to develop a lexicon with stereotypical male and female words. The research project described in this article is currently in the evaluation phase of the final CRISP-DM iteration of the first DSR cycle. We plan to perform further iterations to improve the lexicon, as shown in Figure 1. The dashboard design also evolved iteratively, with a six-person interdisciplinary team of information systems researchers, economists, and psychologists discussing which information should be displayed and what visualization techniques should be used. In total, the dashboard development underwent five iterations. After each iteration, we asked a team of six researchers not involved in the research project for feedback. This feedback was implemented accordingly in the subsequent iteration. To evaluate the first prototype, we conducted an interim measurement of an artificial formative evaluation using an online experiment (Venable et al. 2016). At the beginning of the experiment, we explained to participants how we developed the GLA. Then participants rated how much they agreed with the GLA results of eight sample texts. For four of the eight texts, we asked participants to rate the gender bias score of the text themselves before seeing the results so that we could compare human assessment with the automated assessment of the GLA. Finally, we used well-established scales to measure ease of use (Dolan et al. 2013) and reactance (Ehrenbrink and Möller 2018) regarding the GLA. In addition, participants indicated their affinity for technology (Franke et al. 2019) and whether they perceive themselves as victims of discrimination due to their gender (Kobrynowicz and Branscombe 1997). These variables achieved Cronbach's alpha values of 0.82 to 0.94; thus, reliable results can be assumed. In addition to the quantitative measurements, we included open-ended questions to enrich and refine the meta-requirements and design objectives.

Deriving Meta-Requirements and Design Objectives

We gathered information from the expert interviews and literature to identify the requirements. Thereafter, we synthesized these requirements into meta-requirements and specified design objectives (Figure 2). **MR1–2** and **DO1** address the gender bias analysis results and provide education on gender discrimination by raising awareness about unconscious bias. Through a transparently and intuitively designed dashboard, we provide information about gender bias in the text and explain which parts of the text have led to it.

MR3–4 and **DO2** relate to how users can apply recommendations provided by the GLA. However, recommendations on using gender-inclusive language can quickly cause reactance. Therefore, recommendations from the GLA must find a balance between concrete recommendations and an unobtrusive tone (Sun et al. 2022). **MR5–6** and **DO3** address the previously mentioned issue that textbook analyses have primarily been performed manually, which is time- and personnel-intensive. To address this problem, the GLA should automate these types of analyses to present extensive results regardless of the size of the text corpus. The problem with some automated lexicon approaches is that they are limited in size and coverage and lose relevance over time (Cryan et al. 2020). This can be solved by combining a lexicon approach with a self-learning system, which we will implement in the future. **MR7** and **DO4** cover a range of different usage scenarios (De Waard and Zolfo 2009). Some use the tool to deconstruct stereotypes, while others want to know if their text is appealing to girls or boys. In some cases, users do not want to achieve gender balance because of their target group but want to bias their text in a certain direction, for example to attract girls to participate in a coding course. This requirement can be met by allowing individual setting options.

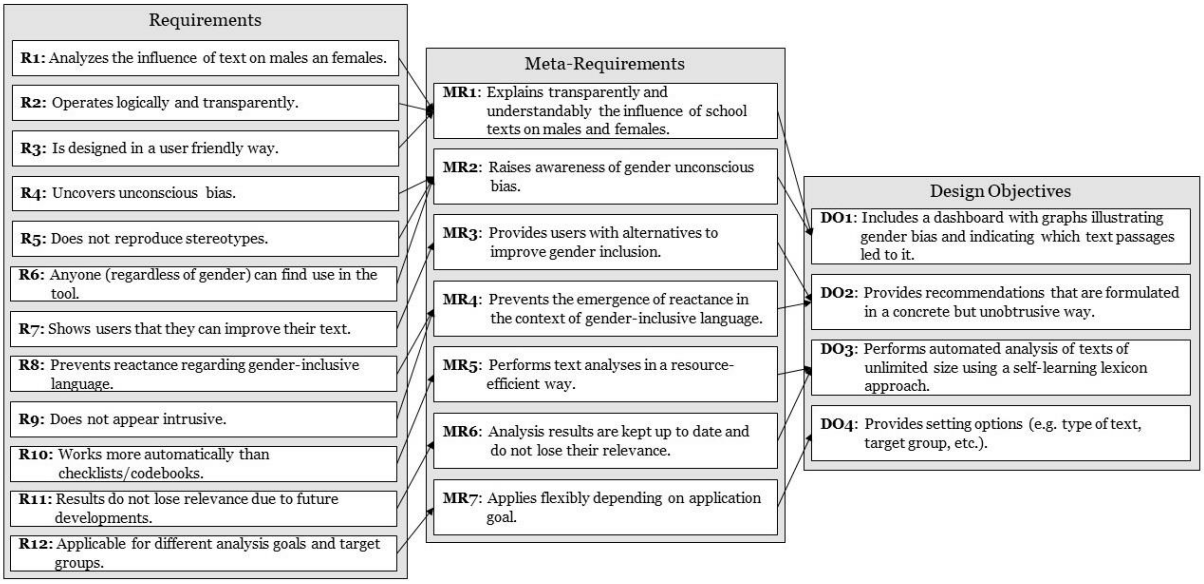

Figure 2. Requirements, Meta-Requirements, and Design Objectives for the Gender Language Analyzer

Development of the Gender Language Analyzer

The following procedure describes the final iteration of the CRISP-DM process. A detailed explanation of each iteration would exceed the scope of this article. Aiming at the evaluation of textbooks and the resulting target group of younger children, we took a different approach to develop a lexicon than previous studies. We used children's books and movie transcripts to ensure the database of the lexicon was close to the evaluated textbooks. Another reason is that gender socialization develops during childhood and is influenced by what children watch and read in media. We gathered information on children's books and movies, particularly regarding how many girls and boys consume them, what the bestsellers are in the boys' and girls' categories, and which are recommended for girls or boys. The information was gathered from Amazon.com, ffa.de and weltbild.de and can be provided on request. We downloaded transcripts of books and films that were specifically recommended for boys or girls or that were mainly consumed by either group. We performed data cleansing using Python for both movies and books. First, we lemmatized the text using *spaCy* and the *en_core_web_lg model*[1], converted it to lowercase and filtered line breaks, special characters, stop words, and names. The words, which remained after the cleaning, were stored in a database

[1] https://spacy.io/

containing information about whether they originated from either girls' or boys' books or movies. Thereafter, we conducted further data preparation using SQL. First, we calculated how often individual words occur in the respective text source. We distinguished between an absolute value and a relative value of how often the word occurs per 1,000,000 words. The distribution of each word per gender across all sources was calculated as a percentage. To avoid an influence of frequently occurring words on the total score, the 10% words with the highest frequency, which are distributed between 40% and 60% for male and female, were removed from the data set. Furthermore, all words that occurred in only one source were filtered from the dataset to avoid too high an influence of individual sources on the overall score. We then further processed the resulting table using Python. We observed that single words, like extreme values, strongly biased the validity of the gender bias analysis. Therefore, we converted the occurrence of the words into an inverted ranking, transforming them from a metric to an ordinal scale level. For example, in a text with 500 unique words, the most frequently occurring word has a rank of 500, and the rarest word has a rank of 1. We calculated the keyness for each word in R with the function *textstat_keyness*.[2] The input comprised two text corpora, one for male and one for female. The text corpora contain all words multiplied by the corresponding ranking value.

We created a data pipeline that determined the displayed values using the lexicon to visualize the analysis results on a dashboard (Figure 3). For this purpose, we performed data cleaning of the textbooks by again using the Python library spaCy with the en_core_web_lg model, as described above. In contrast to the first data processing, only nouns, adjectives, verbs, and proper nouns were considered. Since the appearance of male and female characters is often part of gender bias analyses for textbooks, we further used *spaCy* to recognize character names automatically. Based on the filtered names, the gender of the characters appearing in the text was determined using the genderComputer tool[3] from Vasilescu et al. (2014). Afterward, the names were aggregated further into groups so that a single name was counted only once. Over several iterations, we fixed problems such as characters with the same first name or non-recognition of honorifics such as Mrs. Next, we used the lexicon to calculate the total gender bias score of the text, the total score of all male and female words, the occurrence of male and female words in the text, and the frequency of each word. We subsequently sorted the words by the strength of their influence on the text and generated a top 10 list of the most influential words. Furthermore, to improve understanding, the dashboard displays interpretation aids and improvement suggestions to the user based on the analyzed text.

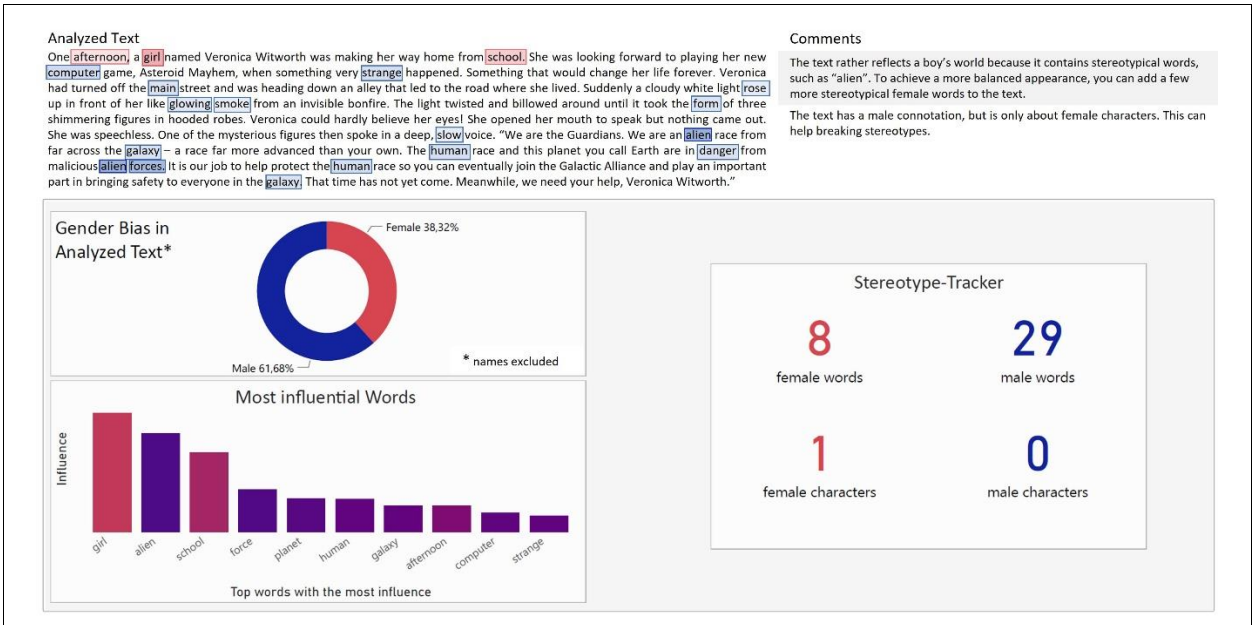

Figure 3. The GLA Dashboard

[2] https://quanteda.io/reference/textstat_keyness

[3] https://github.com/tue-mdse/genderComputer

The dashboard is visualized in Figure 3. The original text can be seen in the upper left corner. Stereotypical words are marked in red for female and blue for male – the intensity of the color indicates whether the word is slightly, fairly, or strongly stereotypical. Words with low scores are not highlighted. Here, stereotypical means words that appear in children's books and movies that are primarily consumed by boys or girls. Thus, in Figure 3, the words "alien," "force," and "galaxy" are highlighted because these are topics that appear more frequently in stereotypical boys' movies (e.g., Star Wars) than girls' movies and consequently influence their gender socialization. In the upper right comment section, the user receives suggestions for combating gender bias in their texts. Each user can decide for themselves whether or not to follow these suggestions. The donut diagram on the left side shows the distribution of stereotypically male and female words and explains that the gender of the characters does not influence the score. In the lower left corner, a bar chart depicts up to 10 most influential words in the text. The color gradient is red for female, blue for male, and purple for a balanced word that tends only slightly toward one direction. Four indices in the lower right corner visualize the total number of stereotypical female and male words and the number of female and male characters.

Results of the Evaluation Experiment

In our study, 33 people participated in the evaluation experiment, of whom 14 identified as male and 19 as female. To evaluate the GLA, we used eight sample texts, ranging from 177 to 240 words. We selected the text passages from English textbooks that are used in schools for students to practice their reading comprehension. Study participants rated whether they agreed with the GLA results of these eight texts on a scale from 1 ("do not agree at all") to 7 ("completely agree"). For the last four texts, we asked participants to rate the gender bias score of the text themselves before seeing the results. For the first four dashboards, participants reported an average agreement of 4.41 (between the response options "4 – neither agree nor disagree" and "5 – somewhat agree"). For the last four dashboards, agreement was higher at 5.05. This difference was significant according to the paired *t*-test (t = -3.2641, p = 0.003, n = 32) and indicated a medium-size effect (d = 0.57). One explanation for this is that participants became aware that their own assessment did not deviate far from the gender bias score shown in the dashboard after assessing the text's gender bias score themselves (Figure 4). Another finding from the evaluation was that people who subjectively agreed more with the GLA dashboard also rated its ease of use higher (r = 0.70**) and showed less reactance toward it (r = -0.53*). However, we observed no significant correlations between subjective agreement and affinity for technology, nor between subjective agreement and one's perception of being a victim of discrimination, indicating that the GLA is not expected to be adopted only by a specific technology-affine or discrimination-sensitive user group.

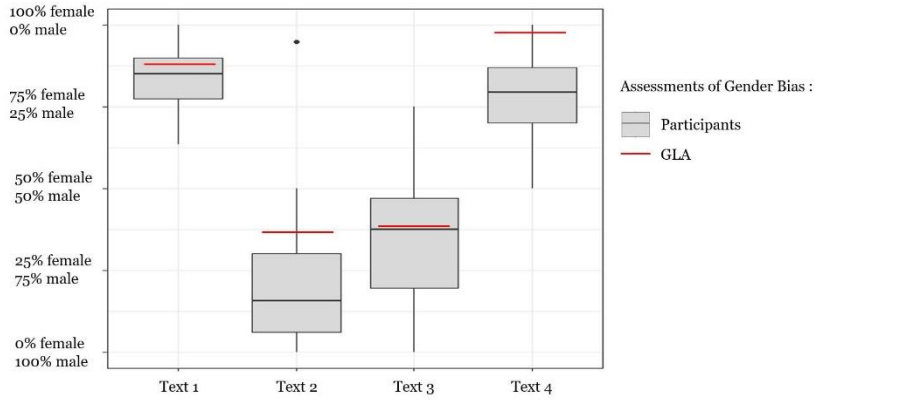

Figure 4. Assessment of Gender Bias in Four Sample Texts

We also asked for feedback via open-ended questions and categorized the answers into five clusters. Cluster 1 relates to the comprehensibility of the results presented in the dashboard. For example, the participants criticized contradictory comments and graphs or unclear marking of certain words as male or female. In addition, participants requested that the GLA considers further contextual information from the text (Cluster 2). For example, they stated that the GLA should better identify whether stereotypical roles are being portrayed in the text. Answers in Cluster 3 revealed criticism of the occurrence counts of male and

female characters. The criticism focused on the accuracy of the function and that the occurrence of male and female characters should impact the overall gender bias score. Participants also suggest that the GLA should specifically identify the gender of the protagonist. Cluster 4 relates to the interface design (e.g., using the stereotypical colors red and blue). Finally, participants suggested that the tool should also consider non-binary genders (Cluster 5).

Discussion

The evaluation shows that our GLA can analyzes texts similarly to humans and recognize gender bias therein. In its current version, our tool is suitable for analyzing texts written for children. It is particularly suitable if an assessment by humans either cannot be done objectively enough or would require considerable resources because of its scope. We found no correlation between agreement with the GLA results and respondents' affinity for technology, which indicates that users do not need high technical knowledge. This enables teachers, publishers, and authors without technical knowledge to use the tool and perform gender bias analyses on a large scale. However, the tool aims not only to analyze the influence of texts on males and females but also to raise awareness of gender bias and support authors in checking their own work.

In the evaluation, our approach has caused criticism regarding the analysis focusing only on males and females. Non-binary people were not considered, and gender-neutral words are inadequately represented. While the inclusion of gender-neutral words in the dashboard tends to be easy to implement through the lexicon, the inclusion of non-binary people is challenging. This problem arises due to the development of the lexicon based on statistical data of consumers of children's books and movies. Non-binary people are not represented in these statistics. Therefore, we will seek alternative data sources for the next research cycle to allow such an inclusion. For the second cycle of the DSR process (in-depth analysis), we aim to use further natural language processing techniques to analyze context, thereby ensuring a more in-depth analysis compared with the first cycle. For example, we will use named entity recognition to identify occupational titles and thus make statements about whether the profession in which a person is portrayed is stereotypical. In the evaluation, participants criticized the GLA's contradictory suggestions for improvement. Future evaluation cycles will improve the quality of the suggestions and address the question of how users should handle them. Moreover, the evaluation results indicate that participants did not fully understand some markings of words as male or female. This is due to minor biases in the lexicon caused by the small training dataset. As a result, each script or book has a substantial impact on the lexicon in the current version. We will improve this in future versions by expanding the word base. We also plan to add more adult vocabulary to the lexicon to be able to apply the GLA to diverse texts flexibly. Based on the success of the previous approach, we will again use media that is consumed by people who identify as male, female, or non-binary. According to current planning, the third and final cycle (self-learning system) will address the criticized lexicon approach (Cryan et al. 2020). We will implement a self-learning system by using online sources that are constantly expanding. So far, we have considered two possibilities. First, we could find sources that classify words explicitly in terms of individuals who identify as either male, female, or non-binary. Second, by this stage, the analysis of texts should have improved to the point where the algorithm can reliably identify the group to which the text refers. For example, large text datasets from Twitter or Reddit can be analyzed and used to expand the database. These two approaches will be compared in the CRISP-DM iterations of the third cycle. Each cycle will conclude with an interim evaluation to formatively evaluate the artifact and compare different prototype versions. In the third cycle, we will further compare our lexicon with existing lexica from research and practice. We are optimistic that the self-learning system will detect changes in the perception of stereotypes and thus map them into the data, especially since we expect it will be a slowly changing process reflected in the data over time.

Conclusion

In our paper, we demonstrate that automated detection of gender biases based on a lexical approach is possible even without prior categorization of words by humans. Our tool is able to analyze texts similarly to humans and thus offers the possibility to support human assessments, especially if the analysis of a large amount of text is necessary (e.g., from many textbooks). We did not identify a significant correlation between affinity for technology and agreement with the GLA. This aspect is of great importance for the tool, as it indicates that the tool can also be used by people who are unfamiliar with information systems and

computer science (e.g., some primary education teachers). Nevertheless, extension and improvement of the tool based on the feedback from respondents summarized in this paper are necessary. Therefore, we will follow the iterations of the DSR cycles to improve the tool so that it can, on the one hand, be made available to the public and, on the other hand, retain its relevance over time, unlike previous lexical approaches. By identifying requirements, meta-requirements, and design objectives, we enrich the body of knowledge in information systems and provide a starting point for further research. In the following research cycles, we will continue to improve the tool and synthesize design knowledge. This design knowledge can then serve as a foundation for other applications, e.g. for the identification and reduction of bias. This will benefit both researchers and companies.

Based on our results, we are optimistic that in the future, the GLA will be able to contribute to the automated analysis of texts for gender bias even outside of the educational domain. In other domains, the efficient analysis of large amounts of text may also be more important, for instance due to rapidly changing texts (e.g., in social media agencies). Furthermore, we envisage that our tool will be able to use image recognition techniques to identify visual gender bias. Our approach may even be transferable to other aspects of subtle discrimination, such as bias based on age or race.

Acknowledgements

The authors would like to thank the interview and evaluation participants, the project team Laura Hein, Dana Dix, Berna Eybey, Enrico Kochon, Sergey Krutikov, Jan Schulte to Brinke, and Katharina Illgen for their valuable and substantive help, as well as the reviewers for their constructive feedback.

References

Bem, S. L. 1974. "The Measurement of Psychological Androgyny," *Journal of Consulting and Clinical Psychology* (42:2), pp. 155–162. (https://doi.org/10.1037/h0036215).

Blakemore, J. E. O., and Centers, R. E. 2005. "Characteristics of Boys' and Girls' Toys," *Sex Roles* (53:9), Springer, pp. 619–633.

Cheryan, S., Meltzoff, A. N., and Kim, S. 2011. "Classrooms Matter: The Design of Virtual Classrooms Influences Gender Disparities in Computer Science Classes," *Computers and Education* (57:2), Elsevier Ltd, pp. 1825–1835. (https://doi.org/10.1016/j.compedu.2011.02.004).

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., and Zhao, B. Y. 2020. "Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–11. (https://doi.org/10.1145/3313831.3376488).

Dele-Ajayi, O., Bradnum, J., Prickett, T., Strachan, R., Alufa, F., and Ayodele, V. 2020. "Tackling Gender Stereotypes in STEM Educational Resources," *Proceedings - Frontiers in Education Conference, FIE* (2020-Octob). (https://doi.org/10.1109/FIE44824.2020.9274158).

Diederich, S., Brendel, A. B., and Kolbe, L. M. 2020. "Designing Anthropomorphic Enterprise Conversational Agents," *Business and Information Systems Engineering* (62:3), Springer Fachmedien Wiesbaden, pp. 193–209. (https://doi.org/10.1007/s12599-020-00639-y).

Dolan, J. G., Veazie, P. J., and Russ, A. J. 2013. "Development and Initial Evaluation of a Treatment Decision Dashboard," *BMC Medical Informatics and Decision Making* (13:1). (https://doi.org/10.1186/1472-6947-13-51).

Ehrenbrink, P., and Möller, S. 2018. "Development of a Reactance Scale for Human–Computer Interaction," *Quality and User Experience* (3:1), Springer International Publishing, pp. 1–13. (https://doi.org/10.1007/s41233-018-0016-y).

Elgar, A. G. 2004. "Science Textbooks for Lower Secondary Schools in Brunei: Issues of Gender Equity," *International Journal of Science Education* (26:7), pp. 875–894. (https://doi.org/10.1080/0950069032000138888).

Fast, E., Vachovsky, T., and Bernstein, M. S. 2016. "Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community," *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016* (Icwsm), pp. 112–120.

Franke, T., Attig, C., and Wessel, D. 2019. "A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale," *International Journal of Human-Computer Interaction* (35:6), pp. 456–467. (https://doi.org/10.1080/10447318.2018.1456150).

Gaucher, D., Friesen, J., and Kay, A. C. 2011. "Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality," *Journal of Personality and Social Psychology* (101:1), pp. 109–128. (https://doi.org/10.1037/a0022530).

Hannover, B., and Kessels, U. 2004. "Self-to-Prototype Matching as a Strategy for Making Academic Choices. Why High School Students Do Not like Math and Science," *Learning and Instruction* (14:1), pp. 51–67. (https://doi.org/10.1016/j.learninstruc.2003.10.002).

Kerger, S., Martin, R., and Brunner, M. 2011. "How Can We Enhance Girls' Interest in Scientific Topics?," *British Journal of Educational Psychology* (81:4), pp. 606–628. (https://doi.org/10.1111/j.2044-8279.2011.02019.x).

Kerkhoven, A. H., Russo, P., Land-Zandstra, A. M., Saxena, A., and Rodenburg, F. J. 2016. "Gender Stereotypes in Science Education Resources: A Visual Content Analysis," *PLoS ONE* (11:11), pp. 1–13. (https://doi.org/10.1371/journal.pone.0165037).

Kobrynowicz, D., and Branscombe, N. R. 1997. "Who Considers Themselves Victims of Discrimination? Individual Difference Predictors of Perceived Gender Discrimination in Women and Men," *Psychology of Women Quarterly* (21:3), pp. 347–363. (https://doi.org/10.1111/j.1471-6402.1997.tb00118.x).

Kuechler, B., and Vaishnavi, V. 2008. "On Theory Development in Design Science Research: Anatomy of a Research Project," *European Journal of Information Systems* (17:5), pp. 489–504. (https://doi.org/10.1057/ejis.2008.40).

Lee, J. F. K., and Collins, P. 2009. "Australian English-Language Textbooks: The Gender Issues," *Gender and Education* (21:4), pp. 353–370. (https://doi.org/10.1080/09540250802392257).

Madera, J. M., Hebl, M. R., and Martin, R. C. 2009. "Gender and Letters of Recommendation for Academia: Agentic and Communal Differences," *Journal of Applied Psychology* (94:6), pp. 1591–1599. (https://doi.org/10.1037/a0016539).

Mayring, P. 2010. *Qualitative Inhaltsanalyse. Grundlagen Und Techniken*, Belz.

McKown, C., and Weinstein, R. S. 2003. "The Development and Consequences of Stereotype Consciousness in Middle Childhood," *Child Development* (74:2), pp. 498–515. (https://doi.org/10.1111/1467-8624.7402012).

Miller, D. 2010. *Stuff*, Cambridge: Polity.

Moser, F., and Hannover, B. 2014. "How Gender Fair Are German Schoolbooks in the Twenty-First Century? An Analysis of Language and Illustrations in Schoolbooks for Mathematics and German," *European Journal of Psychology of Education* (29:3), pp. 387–407. (https://doi.org/10.1007/s10212-013-0204-3).

Parkin, C., and Mackenzie, S. 2017. "Is There Gender Bias in Key Stage 3 Textbooks?: Content Analysis Using the Gender Bias 14 (GB14) Measurement Tool," *Advanced Journal of Professional Practice* (1:1), pp. 23–40.

Ramakrishna, A., Martínez, V. R., Malandrakis, N., Singla, K., and Narayanan, S. 2017. "Linguistic Analysis of Differences in Portrayal of Movie Characters," *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (1), pp. 1669–1678. (https://doi.org/10.18653/v1/P17-1153).

Sun, J., Wu, T., Jiang, Y., Awalegaonkar, R., Lin, X. V., and Yang, D. 2022. "Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages," *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA* (Vol. 1), Association for Computing Machinery. (https://doi.org/10.1145/3491102.3502114).

Vasilescu, B., Serebrenik, A., Devanbu, P., and Filkov, V. 2014. "How Social Q&A Sites Are Changing Knowledge Sharing in Open Source Software Communities," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 342–354.

Venable, J., Pries-Heje, J., and Baskerville, R. 2016. "FEDS: A Framework for Evaluation in Design Science Research," *European Journal of Information Systems* (25:1), pp. 77–89. (https://doi.org/10.1057/ejis.2014.36).

De Waard, I., and Zolfo, M. 2009. "Integrating Gender and Ethnicity in Mobile Courses Ante-Design: A TELearning Instrument," *International Journal of Interactive Mobile Technologies (IJIM)* (3:1), p. 77. (https://doi.org/10.3991/ijim.v3i1.674).

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. 2015. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp. 454–463