University of Cologne
Faculty of Arts and Humanities
Department for Digital Humanities

Seminar "Embeddings"
Summer Term 2020

Seminar Work

A Comparative Analysis of Implicit Biases in Word Embeddings Trained on 19th and 20th Century Children's Literature

by
Viktoriya Olari

Seminar Work Submitted in Partial Fulfilment of the Requirements for the Degree of
Master of Arts in Information Processing

Seminar Work Supervisor: Dr Nils Reiter
January 25, 2021
Cologne

Abstract

Previous research has shown that children's literature reinforces the process of forming prejudices in children from an early age. However, while prior studies have mostly focused on identifying bias in selected books published in the 1960s and later, there has been little research that analyses large volumes of children's literature published in the 20th century and earlier. In this work, the author fills this gap by examining biases inherent in the language of children's books through an analysis of large body of 19th and 20th century children's literature. To do so, the author developed a Bias Assessment Module which an application that measures biases in sets of word embeddings in the corpus of children's literature using the Word Embeddings Association Test (WEAT). The WEAT is a state-of-the-art approach that is used to assess the occurrence of known biases in large volumes of text. Additionally, the author implemented a method for discovering new biased word sets by clustering the word embeddings, and applied this method to one of the corpora with sparse vocabulary. The findings indicate that word embeddings in children's literature throughout the centuries exhibit gender, religious, and racial biases. Stereotypes about animals, who are often used as main characters in children's books, could not be clearly confirmed based on the results. Although the clustering of word embeddings returned new sets of biased words, the word selection in each of the sets seemed to be random and to lack cohesion. Future research could focus on applying further unsupervised methods to find new biased word sets in word embeddings trained on children's books and child-specific delineation of words used for the WEATs. Such research could also investigate larger corpora of children's books and employ additional techniques to assess bias.

Table of Contents

List of Tables ..... **v**

List of Figures ..... **vi**

List of Abbreviations ..... **vii**

1 Introduction **1**

1.1 Methodology Overview and Research Questions 1

1.2 Structure 2

2 Background and Related Work **2**

2.1 Investigating Prejudice in Children's Literature 2

2.2 Detecting Bias in Word Embeddings with the WEAT 4

2.2.1 An Introduction to the WEAT's Terminology 4

2.2.2 Use Cases of the WEAT 4

2.3 Limitations of the WEAT and Unsupervised Approaches for Discovering the Target Words 5

2.4 Summary and Analysis of Shortcomings 6

3 Methodology **6**

3.1 Preparation of the Corpora and Training of the Word Embeddings 7

3.1.1 Choice of Children's Books Corpora 7

3.1.2 Pre-processing of the Corpora 8

3.1.3 Training of the Word Embeddings 8

3.2 Experimental Setup 8

3.2.1 WEAT: An Experimental Protocol 8

3.2.2 Definition of Attribute and Target Words for the WEAT Experiments 9

3.2.3 Identifying New Sets of Target Words 10

3.3 Testing Procedure 11

3.4 Limitations 12

4 Bias Assessment Module **12**

4.1 System Architecture 12

4.1.1 File Structure 12

4.1.2 Class Structure 13

4.2 Implementation Details 14

4.2.1 Module Initialisation 14

4.2.2 Corpus Processing and Supplying the Models 15

4.2.3 Execution of the WEAT Experimental Protocol 15

4.2.4 Finding New Target Words for Specific Bias Categories with Clustering 16

4.2.5 Logging the Test Results 16

5 Results **16**

5.1 Replicating Results from Previous Studies 17

5.2 WEAT Results for 19th and 20th Century's Children's Literature 17

5.2.1 Gender Bias 17

5.2.2 Religious Bias 18

5.2.3 Animal Stereotypes 18

5.2.4 Age-related and Racial Biases 18

5.3 Results for Detecting New Bias Word Categories in CPB Corpus 18

6 Discussion **19**

6.1 Summary of and Reflections on the Results 19
6.2 Evaluation of the Study Results 20
7 Conclusion 21
8 Bibliography 23
9 Appendix 26
9.1 Table with Target and Attribute Words 26
9.2 Sequence Diagram of the Bias Assessment Module 28
9.3 Table with the WEAT results 29
9.4 Table with the Out-Of-Vocabulary Words for each WEAT 31
9.5 Changelog 32
10 Statement of Independent Work 43

List of Tables

Table 1: An overview of the corpora used to train the word embeddings. 7
Table 2: Overview of the categories to be tested with the Bias Assessment Module and the sources from which the attribute and target words were derived. 10
Table 3: List of arguments that the user can input on start of the Bias Assessment Module. 14
Table 4: Configuration objects in `BiasAssessmentModule` 14
Table 5: File names and contents for the storing of the WEAT results. 16
Table 6: WEAT results for GoogleNews and GAP corpus for G1–G5 categories obtained in the current study (on the left) in comparison to the original findings (on the right) published by Chaloner and Maldonado (2019b). The p-values in bold indicate statistically significant bias, $p < 0.05$. 17
Table 7: Selection of induced CG1, CG2 and CA1 biased word categories per cluster (an extract). .. 19
Table 8: Target and attribute word lists used for the WEAT. 26
Table 9: WEAT results, p-values in bold indicate statistically significant gender bias $p < 0.05$. 29
Table 10: Number of out-of-vocabulary target and attribute words in the ChiLit, CLLIP and CPB embeddings. 31

List of Figures

Figure 1: Class architecture of the Bias Assessment Module. 13
Figure 2: Simplified sequence diagram of the Bias Assessment Module for the example of supplying the ChiLit corpus. 28

List of Abbreviations

AW	attribute words
BNC	British National Corpus
b	billion
CBOW	continuous bag of words
ChiLit	19th Century Children's Literature Corpus
CLLIP	Corpus-based Learning about Language In the Primary-school
CPB	Children's Picture-Book Corpus
c.	century
GLoVe	Global Vectors
IAT	Implicit Association Test
m	million
POS	Part of Speech
TW	target words
WEAT	Word Embedding Association Test

1 Introduction

Children are often influenced by societal norms and exhibit implicit biases[1] towards others from an early age (Cole and Valentine 2000; Baron and Banaji 2006; Raabe and Beelmann 2011; Cvencek, Meltzoff, and Kapur 2014; Babcock et al. 2016). Leahy and Foley (2018) and McCabe et al. (2011) meta-analyses conclude that among other possible sources of biases, such as a child's social environment, children's literature – here, meaning books published for an audience comprised of children and teenagers up to 19 years old (Eberhardt 2018) – reinforces the process of establishing prejudices.

Prejudices in children's literature have been extensively studied. However, previous research has mostly focused on identifying biases in children's books published in the 1960s or later. Prior studies analysed pictorial (Anderson 2013; Turner-Bowker 1996) and textual content (Darni 2017; McArthur and Eisen 1976), but only in selected books or book series. There is no research that studies the language used in children's books through an analysis of a large sample of 19th and 20th century children's literature.

In this seminar paper, the author fills this research gap by examining children's literature for biases inherent in language by means of word embeddings – an approach for representation of words in their textual context as vectors in a high-dimensional space (Caliskan, Bryson, and Narayanan 2017). The following corpora were chosen as being representative of children's literature from the 19th and 20th centuries:

1) The 19th century Children's Literature corpus (ChiLit corpus), which includes a total of 4.4 million words (Čermáková 2017),
2) The extended Corpus-based Learning about Language in Primary School (CLLIP corpus), which is part of the British National Corpus and comprises texts written in the 20th century for child and teenage audiences, comprising a total of 2.5 million words (Thompson and Sealey 2007), and
3) The Children's Picture-Book Corpus (CPB corpus), which contains a total of 65,008 words in 100 children's picture books published in the 20th century (Montag, Jones, and Smith 2015).

In order to measure biases in word embeddings trained on children's literature, the author used the Word Embedding Association Test (WEAT). The WEAT is a state-of-the-art test that has been widely utilised to assess the occurrence of known biases in word embeddings trained on large volumes of text (Caliskan, Bryson, and Narayanan 2016; Tan and Celis 2019; Kurita et al. 2019; Babaeianjelodar et al. 2020; Chaloner and Maldonado 2019b).

1.1 Methodology Overview and Research Questions

Since there are no prior studies that measure biases in word embeddings trained on children's literature that use the WEAT, the author developed the theoretical framework using previous studies that have investigated biases in children's books and have used the WEAT to assess biases in word embeddings. The author first determined which prejudices are detectable in children's book corpora. Afterward, the author explored previous cases where the WEAT was used. Considering the result, the author evaluated whether the WEAT is a feasible method to measure the biases identified in previous research in word embeddings trained on children's literature. Since the WEAT has limitations, the author also explored possible solutions for identifying new categories of biased words.

Based on the findings and shortcomings of previous research, the author developed a Bias Assessment Module in the practical section of this work. She used the module to train the sets of word embeddings on children's books and to measure biases hidden in these word embeddings with the WEAT. She subsequently identified new biased word sets by clustering the word embeddings of CPB corpus with the fewest number of words. Finally, the author compared the biases measured in ChiLit

[1] Bias refers to prior information which is derived from precedents that are known to be harmful. Implicit bias refers to unconscious bias, whereby the person is unaware they are biased. In this work, implicit bias means a bias hidden in the use of language in children's books. Harmful biases can also be called prejudice (Caliskan, Bryson, and Narayanan 2017). The terms 'prejudice' and 'bias' are used interchangeably in this seminar work.

with those measured in the CLLIP and CPB corpora, as representatives of two different time periods and discussed new categories of biased words that were detected.

In summary, the author answers three research questions in this seminar paper:

1. To what extent can known implicit biases be found in word embeddings trained on corpora consisting of children's books?
2. What new biased word sets can be identified?
3. What discrepancies in the significance of prejudice can be found when comparing sets of word embeddings trained on the 19th century ChiLit corpus and 20th century CLLIP and CPB corpora?

1.2 Structure

Chapter 2 outlines the theoretical framework for the practical part of this seminar work. The author substantiates her research goal by investigating known prejudices in children's literature. Based on previous studies that have used the WEAT, the author assesses the feasibility of measuring the biases identified with the WEAT and investigates solutions to overcome the WEAT's limitations. The chapter closes with a summary of findings and shortcomings.

Chapter 3 presents the methodology from three perspectives. First, chapter details the preparation of the corpora and training of the word embeddings. The author explains the selection of the corpora and outlines the procedure chosen to train the single sets of word embeddings. Second, the experimental setup, which serves as a basis for the development of the Bias Assessment Module, is described. The author then presents the WEAT algorithm, the procedure to define the target and attribute words for the WEATs and the algorithm for identifying new sets of biased words. In the third step, the testing procedure is discussed.

Chapter 4 provides an overview of the design and technical implementation of the Bias Assessment Module. The author explains the system architecture of the module and outlines the implementation details.

Chapter 5 reveals the results for measuring biases in word embeddings trained on 19th and 20th century children's literature using the Bias Assessment Module. The author first verified that the WEAT algorithm was performing correctly by comparing the outcome of the WEAT with the results from previous studies that had used the same WEAT algorithm. The chapter then presents the results of the WEAT for the single bias categories that were summarised in Section 3.2.2 and details the discovery of new bias word categories in the CPB corpus.

In Chapter 6, the results of this seminar work are discussed and placed into the context of previous research. The author compares the outcome of the WEATs for 19th and 20th century children's literature and discusses factors that may have influenced the measurements. Chapter 7 lists the main research findings, providing answers to the research questions posed in Section 1.1. Finally, the author reflects on the entire study and makes suggestions for the future research.

2 Background and Related Work

Scholars have studied prejudice in children's literature extensively in several case studies. However, there is a lack of research examining prejudice in word embeddings trained on children's literature. Thus, the following chapter discusses this body of scholarship from three perspectives. First, in order to identify possible bias categories in children's books published in the 19th and 20th centuries, the author reviews related studies that have investigated prejudice in children's literature for this period in Section 2.1. Second, in Section 2.2, the author assesses the feasibility of measuring identified biases in word embeddings. She investigates the existence of studies that have used WEAT to measure prejudice in bias categories identified in Section 2.1. Third, since the WEAT has limitations, the author explores possible solutions for identifying new categories of biased words in Section 2.32.1. Finally, the author summarises the findings and analyses shortcomings of related studies.

2.1 Investigating Prejudice in Children's Literature

Beginning in the 1930s, researchers have become increasingly focused on investigating biases in children's books (Child, Potter, and Levine 1946; Turner-Bowker 1996). However, while many studies have focused on exploring prejudice in children's literature, many of them analysed pictorial content to identify the presence of gender, racial and age-related biases. Turner-Bowker (1996) criticises

children's literature between 1980s–1990s as 'not presenting to our children an accurate picture of contemporary life' (p. 477) and states that female characters are presented in pictures significantly less often than male characters. Cole and Valentine's (2000) work reflects previous studies on how interracial topics are portrayed, arguing that multi-ethnic characters are rare in children's books published between 1970–1990. Anderson (2013) analyses the presentation of male and female characters in science-oriented children's books. The results show that children's books exhibit severe gender bias in both the ratio of male to female images and in the stereotyping of character roles. Crawford (2000) stresses that older adults are stereotyped in the images in children's books; a disproportionate number of grandfathers are depicted as being bald or grey-haired, and disproportionate number of grandmothers are represented as rocking chair-bound women who frequently wear aprons.

Several researchers have investigated the language used in single books and small book collections. Most scholarship of this sort concentrates on linguistic representations of gender biases in children's books. Heywood (2020) points out that gender prejudices became a regular feature in the children's literature of 19th century due to the emergence of strict demarcation of gender roles. Child, Potter, and Levine (1946) analysed children's text-only books published in the 1930s and 1940s. They find that the treatment of female characters closely follows stereotypes for the upbringing of girls that were commonplace in the 19th century. Female characters take what is offered, and they wish, ask, and nurture, whereas male characters are autonomous, dominant, and aggressive. McCabe et al. (2011) conducted a meta-analysis of the titles and central characters of 5,618 children's books published in the United States in the 20th century and found that female characters are significantly underrepresented across all decades and in all book series. Turner-Bowker (1996) analysed the adjectives that the authors of children's books published in the 1980s–1990s chose to describe boys and girls. Her findings suggest that the adjectives the authors used are consistent with gender stereotypes – male characters were described as being active and powerful more often than women were.

Few studies have investigated the racial and religious biases present in texts of children's books. MacCann (2013) analysed characterisations of African Americans in children's books published between 1830–1900. She stated that with few exceptions, African American identities were presented as being less valued than European American ones: 'With the exception of a few abolitionist[2] narratives, children's books have generally treated Black characters stereotypically, or they have excluded them entirely'. Long (1984) analysed children's books published in the 1970s and 1980s and reports that authors largely ignored or mishandled interracial topics in children's literature. Morgan and Forest (2016) examined children's books published during World War II in Germany, and showed that they strongly encouraged anti-Semitic thought. Eberhardt (2018) summarises that even in modern children's books published in Europe, religions other than Christianity are depictured to be foreign, and the Islamic way of live is often criticised. However, Ghesquière (2005), Hunt (2005), Trousdale (2011) assert that due to advancing secularisation, the impact of religion on education and children's literature declined during the final decades of the 19th century. Further, they note that religion had almost disappeared from children's books by the second half of the 20th century.

Several researchers have examined age-related bias in texts of children's books published in 20th century. For instance, Taylor's (1980) paper reflects previous research on stereotypic attitudes against elderly people. Taylor states that the elderly are often described as more passive, less healthy, poorer at problem-solving, and less emotional and self-sufficient than other adults. McGuire (2016) claims that the heterogeneity of aging is underrepresented in children's books. Most children's books portray older characters as grandparents, while in reality, older people have various roles in life and society.

During the 19th and 20th centuries, children's books featuring animals as central characters were popular (Wehrmeyer 2010; Child, Potter, and Levine 1946; Burke and Copenhaver 2004). Azmiry (2014) suggests that authors of children's books use animals to eliminate racial stereotypes. However, research has shown that even stories about animals often exhibit biases in two ways. First, these stories may incorporate biases that are inherent in humans (Dunn 2011). For instance, McCabe et al. (2011) find that throughout children's literature written in the 20th century, most animal characters are sexed, and inequality is greater among animals than humans. They also discover that male animals are featured more often than female animals. Derman-Sparks, Goins, and Edwards (2020) point out that animal

[2] Abolitionism is a movement in Western Europe and the Americas to end the slave trade and chattel slavery (The Editors of Encyclopaedia Britannica 2019).

families in children's books are often described as including two parents and living in isolation, which is almost always incorrect in real life. Second, stories about animals may incorporate biases and stereotypes that are only particular to animals. For example, Dunn (2011) examines children's literature from the 20th century. The author finds that while dogs are described as being loyal, wolves and foxes are generally characterized as villains. Compared to dogs, domestic cats are ascribed a wide variety of personalities in children's books; they are described as being loyal, but also possessing cunning and manipulating characters (Wehrmeyer 2010). Mice are the most successful animals in children's literature – they are portrayed as small and often overlooked and ignored, but at the same time shown harbouring lively, cute, and secret personalities (Dunn 2011).

2.2 Detecting Bias in Word Embeddings with the WEAT

The studies presented in Section 2.1 demonstrate that children's literature exhibits many types of prejudice, such as biases against certain genders, races, religions and age of a person. Even children's books which feature animals as central characters may suffer from the presence of prejudice. The following section examines whether there are previous studies that have investigated these biases in word embeddings using the WEAT. The author first introduces the reader to the terminology used in the WEAT. Subsequently, she analyses the previous studies on the common use cases of the WEAT.

2.2.1 An Introduction to the WEAT's Terminology

The WEAT borrows terminology from the Implicit Association Test (IAT) used in psychology to account the individual implicit prejudices (PrÈcenth 2019; Kurpicz-Briki 2020). The purpose of the IAT is to measure how strongly a person associates two target concepts (for instance, *programmer* and *nurse*) with an attribute (for instance, *female*) based on the reaction time of the subject person. The higher the subject's latency to respond to a particular pairing of a concept with an attribute (*programmer* + *female*) is, relative to the paring of another concept with the same attribute (*nurse* + *female*), the weaker the association between the concept and the attribute is (Greenwald, McGhee, and Schwartz 1998; Lane et al. 2007).

Instead of using the latency to measure the differential association between a single pair of target concepts and an attribute, the WEAT uses cosine similarity to measure the differential association between two sets of target concepts and an attribute concept represented as vectors in word embeddings (Caliskan, Bryson, and Narayanan 2017). Each set of two target concepts consists of words that the researcher suspects are biased towards the attribute concept, represented by two sets of attribute words (Mulsa and Spanakis 2020; Caliskan, Bryson, and Narayanan 2017). As an example, in the context of gender bias associated with specific occupations, target words can be represented by terms like *programmer, scientist,* and *nurse, librarian*, while the attribute concept, 'gender', can be represented by words such as *he, men, male* and *she, women, female.*

While the WEAT algorithm is described in detail in Section 3.2.1, it is necessary to state at this point that bias will exist if words *nurse* and *librarian* are more likely than words such as *programmer* and *scientist* to be closer to terms *she, women, female* than to *he, men, male*. In other words, the bias exists, if women are more associated with the occupations such as nurse and librarian, whereby the men are more associated with occupations such as programmer or scientist.

2.2.2 Use Cases of the WEAT

Since its introduction by Caliskan, Bryson, and Narayanan (2017), the WEAT has become the most common statistical test to measure biases in word embeddings (Ethayarajh, Duvenaud, and Hirst 2019; Popovic, Lemmerich, and Strohmaier 2020). In the following section, the author introduces studies that have used the WEAT to investigate gender, racial, religious, age-related and animal-related biases which were identified in children's literature in Section 2.1.

Multiple studies have utilised the WEAT to measure gender bias in word embeddings. Caliskan, Bryson, and Narayanan (2017) performed the WEAT on word embeddings trained on the GoogleNews corpus. They demonstrate that females are associated more strongly with the target concepts, family and arts, while males are more strongly associated with concepts such as career, science, and mathematics. Chaloner and Maldonado (2019b) confirm these results and additionally observe that women are more likely to be related with concepts like appearance and weakness, whereby the men are more likely to be associated with concepts such as intelligence and strength. They also detected the presence of gender biases in word embeddings trained on texts from specific domains, such as social media and biomedical literature. Kurpicz-Briki (2020) applied the WEAT to German and French word

embeddings trained on the Common Crawl and Wikipedia articles. The author demonstrates that women are more closely associated with target words that describe the concept of family than the concept of career. In comparison, Kurpicz-Briki's results show men are more strongly associated with the concept of career than with the concept of family. Mulsa and Spanakis (2020) confirm the presence of gender biases related to the concept of career in word embeddings trained on the Dutch texts.

Discovering racial biases in word embeddings using the WEAT has been the focus of several studies. Caliskan, Bryson, and Narayanan (2017) demonstrate that European American names are more likely to be associated with adjectives describing pleasant attributes than African American names. Rice, Rhodes, and Nteta (2019) build upon these results in their investigation of racial biases entrenched in legal texts. Their findings suggest that African American names are associated much more closely with negative concepts, while European American names are much more frequently related to positive concepts. Barman, Awekar, and Kothari (2019) find similar results in their analysis of racial biases in word embeddings trained on song lyrics.

Additionally, other studies have investigated social constructs using the WEAT, such as preferences for flowers over insects or musical instruments over guns (Caliskan, Bryson, and Narayanan 2017). Few scholars have measured religious biases (e.g., Christianity versus Islam) in word embeddings with the WEAT, such as in the work of Popovic, Lemmerich, and Strohmaier (2020). Barman, Awekar, and Kothari (2019) used the WEAT to analyse whether young people are more likely than old people to be associated with pleasant than with unpleasant concepts; they reported the medium effect size without detailing the probability. Mulsa and Spanakis (2020) investigated age-related bias in Dutch word embeddings and could not confirm the preference of terms describing young people over those describing old people.

2.3 Limitations of the WEAT and Unsupervised Approaches for Discovering the Target Words

Since domain-specific corpora are often small in size, the target words may be absent in the vocabulary of the word embedding or used with insufficient frequency to enable analysis. The reliability of the WEAT with samples that lack target words is therefore difficult to assess (Chaloner and Maldonado 2019b). Additionally, many forms of prejudice are linked to social constructs that may vary depending on the context and cannot be captured fully by using sets of fixed-target words (Swinger et al. 2019). For this reason, a number of unsupervised, data-driven approaches were developed to identify new target words based on the vocabulary available in word embeddings. The following section outlines some of these approaches.

Chaloner and Maldonado (2019b) developed an unsupervised method for discovering new categories of gender-biased words. After clustering the word embeddings using the k-means++ algorithm, they measured the association between each target word from the cluster and the attribute words. Depending on the result, each word in the cluster was placed into either a male- or female-associated set of target words. In order to test whether the sets of target words in each cluster were biased, the authors applied the WEATs on the original word embeddings with the new words found. All WEATs returned significant results.

Swinger et al. (2019) propose the Unsupervised Bias Enumeration algorithm to identify offensive associations of human names with the affiliation to a race and gender in word embeddings. As with the work of Chaloner and Maldonado (2019b), they used the k-means++ algorithm to perform the initial clustering of the word embeddings to identify sets of target words, creating candidates for gender and race bias word categories. Additionally, they clustered the vectors of the attribute words (here, human names). Then, the authors performed a WEAT, computing the statistical confidences by applying the rotational null hypothesis which implemented the random alignment between these cluster sets.

Aran et al. (2020) suggest an unsupervised methodology in order to identify conceptual biases towards gender and religion anchored in word embeddings. They defined an algorithm which finds two sets of the most biased and frequent – or salient – words from the vocabulary in the corpus regarding the attribute concept. For each set of the most salient words, the authors used the k-means algorithm to aggregate the most semantically similar words into two partitions of clusters, based on the distance between embeddings. They then applied the WEAT to each of the clusters from both partitions and the attribute concept. Only the clusters from each partition that returned significant probability towards the

attribute concept were kept. This ensured that the conceptual biases represented in each cluster were relevant and strongly associated with the attribute concept.

2.4 Summary and Analysis of Shortcomings

In following section, the research findings from the Sections 2.1, 2.2 and 2.3 are summarised and analysed. The shortcomings identified serve as a basis for the practical part of this work.

(1) The studies presented in Section 2.1 demonstrate that children's literature from the 19th and 20th centuries exhibits various biases. However, most case studies have focused on selected texts and, as shown in Section 2.2, none of the studies have examined biases in word embeddings trained on children's books.

(2) Research investigating biases inherent in language in 19th century children's books is sparse. The research primarily examines gender and racial biases in selected pieces. These studies show signs of differentiation of language usage based on affiliation with a particular gender or race.

(3) Several studies have been conducted to identify biases in children's books published in 20th century. Many of these studies performed pictorial analyses of gender, racial, and age-related prejudices. Although there are studies that have inspected prejudices in language, most of the research has focused on investigating gender bias only. Religious and racial biases inherent in texts have only been marginally examined. Few qualitative studies have reviewed texts in their entirety, while the others have analysed selected aspects, such as book titles or adjectives that describe the characters. Several qualitative studies have proved that gender, racial, religious, and age-related biases are omnipresent in 20th century children's books.

(4) There are hardly any studies which have comprehensively compared biases in 19th century children's books with the biases in 20th century children's books.

(5) Animals are often the main characters in 19th and 20th century children's books. While some researchers have suggested that animals were used to eliminate stereotypes, other academics claim that the opposite is true; inequality is greater among animal characters than humans. Although some studies have verified human-like biases in descriptions of the animals in 20th century children's books, there is little research on the topic for 19th century works. Only a few studies have investigated the biases that are particular to animal characters in children's literature.

(6) The WEAT has been used to measure various biases in word embeddings trained on corpora from different domains. In several studies, it has confirmed the presence of gender, racial, religious, age-related and animal-related biases in the word embeddings.

(7) No research has been conducted on measuring biases in word embeddings trained on children's literature. However, after examining examples where the WEAT was used, one can argue that measuring gender, racial, religious, age-related and animal-related biases should be feasible for the domain-specific corpora selected in this study.

(8) The WEAT has been criticised for failing when used for domain-specific corpora; some stereotypes are difficult to capture with fixed target words. Therefore, the author analysed which unsupervised methods are useful to determine the target words for gender, racial, and religious bias categories. All of the methods examined build upon the unsupervised k-means clustering algorithm. None of the studies applied such methods on word embeddings trained on children's literature.

3 Methodology

Having analysed the body of research on common biases in children's literature and established the theoretical framework for investigating these biases in word embeddings using the WEAT, the following chapter outlines the methodology for the practical part of this seminar work. First, the author discusses the corpora selected for the study as well as the procedure for the training of the word embeddings in Section 3.1. In Section 3.2, she details the experimental setup which serves as a basis for the implementation of the Bias Assessment Module. Finally, she presents the testing procedure in Section 3.3 and outlines the limitations of this study in Section 3.4.

3.1 Preparation of the Corpora and Training of the Word Embeddings

The following section provides an overview of the selected corpora and the necessary preparatory steps to arrange the corpora in a format suitable for use with the Bias Assessment Module. In the final portion of this section, the author provides details on the procedure for the training of the word embeddings.

3.1.1 Choice of Children's Books Corpora

Table 1 shows the key data relating to the corpora the researcher selected to train the word embeddings in this research.[3] The table outlines the names of each corpus, the type of literature included in the corpus, the estimated age of children to whom the books in the corpus are targeted, the time period covered by the corpus, and the total number of words in each corpus.

Table 1: An overview of the corpora used to train the word embeddings.

Corpus	Type	Age	Time Period	Words
Original corpora				
ChiLit – 19th Century Children's Literature corpus (Čermáková 2017)	Fiction	-	1826–1911	4.4m
Extended CLLIP – Corpus-based Learning about Language in Primary School (Thompson and Sealey 2007)	Fiction, brownies, annuals (short stories, jokes, quizzes)	8–19	1902–2001	2.5m
CPB – Children's Picture-Book corpus (Montag, Jones, and Smith 2015)	Not specified	0–6	1901–2014	65,008
Sub-corpora with reduced vocabulary				
ChiLit Small	Fiction	-	1826–1911	65,008
CLLIP Small	Fiction	8–19	1902–2001	65,008

The ChiLit corpus is representative of children's literature published in the 19th century and is publicly available (Centre for Corpus Research 2019). This corpus includes a sample of works from the Golden Age of English children's literature – fiction written for and read by children in the 19th century. The creators of the corpus did not specify the age of children to whom the books were targeted. The corpus was built to ensure a balanced representation of male and female authors; 35 books were written by female authors and 36 were written by male authors (Čermáková 2017, 2018).

In this study, the 20th century is represented by two corpora: the extended CLLIP corpus and the CPB corpus. The CLLIP corpus is a part of the British National Corpus (BNC Consortium 2007) and is publicly available. Originally, it was comprised of 42 fiction texts written in the 20th century for a child audience (8–10 years old; Thompson and Sealey 2007). Although the creators reduced the number of texts[4] that were already part-of-speech (POS) tagged in the corpus to 30, the author decided to retain all 42 texts, since the POS tagging was not necessary for the purposes of this study. In addition, the author extended the CLLIP corpus by adding 77 books from the BNC that were written for teenage audiences (13–19 years old).

The CPB corpus is not publicly available. The author requested the corpus for this study from Jessica L. Montag, PhD from the University of Illinois, Urbana-Champaign, who created the corpus in 2015. The corpus consists of a representative sample of books that parents read to very young children. The books were selected from lists of librarian-recommended picture books, Amazon.com best sellers, and circulation statistics from the infant and preschool sections of the Monroe County (Indiana) Public Library (Montag, Jones, and Smith 2015).

Since both the ChiLit and CLLIP corpora feature larger vocabularies than the CPB corpus, the author created the ChiLit Small and CLLIP Small corpora in order to investigate how the reduced vocabulary might influence the bias assessment. Each of the corpora contains the same number of words as the CPB corpus.

[3] Today, the largest corpus of children's literature is the Oxford Children's Corpus (Wild, Kilgarriff, and Tugwell 2012). This corpus is continually growing. The books in this corpus are targeted at children aged 5–14 years and contain over 30 million tokens. However, this corpus could not be included in this research, since only research partners of the Oxford University Press have access to it.

[4] Thompson and Sealey (2007) report that their CLLIP Corpus consists of 40 texts. However, when the author queried the BNC, she found 42 texts tagged as being for a child audience.

3.1.2 Pre-processing of the Corpora

While the ChiLit and CPB corpora were in plain text format and could be used directly to train the word embeddings, the CLLIP corpus first had to be extracted from the BNC and transformed into plain text format. To do this, three steps were necessary:

1) In order to identify which texts in the BNC are directed to child and teenage audiences, the author queried BNC using the text analysis software, Sketch Engine (Sketch Engine 2020), with the following statements: `<bncdoc wriaud=="Child">[]` and `<bncdoc wriaud=="Teenager">[]`. The queries returned the IDs of 119 files in XML format, each containing one book.
2) The corresponding files in XML format were identified by ID obtained in 1) and then manually extracted by the author from the full BNC (BNC Consortium 2007).
3) All the tagging and metadata were removed from each file.

The sub-corpora ChiLit Small and CLLIP Small were extracted from the ChiLit and CLLIP corpora respectively. The author designed a mechanism in the Bias Assessment Module that can create multiple unique subsets from any corpus and process them in the Bias Assessment Module as independent corpora. The procedure for specifying the amount of the sub-corpora is explained in detail in Section 4.2.1.

3.1.3 Training of the Word Embeddings

In order to generate a set of word embeddings for each of the corpora, the author used the word2vec algorithm as originally proposed by Caliskan, Bryson, and Narayanan (2017). For the training, the author selected the skip-gram architecture instead of the continuous bag of words (Mikolov, Sutskever, et al. 2013), since it performs better for corpora with infrequent words, such as the domain-specific corpora chosen for this study (Google Code Archive 2013).

When choosing the dimensionality of the vectors, the author drew upon examples of word embeddings used in prior studies (Chaloner and Maldonado 2019b; Caliskan, Bryson, and Narayanan 2017) and trained 300-dimensional word embeddings for each corpus. For each of the corpora, the author used two different sliding window sizes: (1) 10 words, as suggested by the Google Code Archive (2013), and (2) 30 words, since Mikolov, Sutskever, et al. (2013) pointed out that broader contexts can lead to greater accuracy.

3.2 Experimental Setup

The following section describes the experimental setup that serves as the basis for the implementation of the Bias Assessment Module and the subsequent bias measurement. The author first outlines the experimental protocol for measuring bias by formalising the WEAT. Second, she presents the WEAT categories tested with the Bias Assessment Module and describes the procedure for the defining of the target and attribute words. Third, she formalises the unsupervised clustering approach for finding new target words from the embedding vocabulary. Fourth, she describes the testing procedure.

3.2.1 WEAT: An Experimental Protocol

The following section formalises the WEAT hypothesis protocol, whereby the protocol structure was aggregated from several descriptions of the WEAT (Mulsa and Spanakis 2020; Caliskan, Bryson, and Narayanan 2017; Rice, Rhodes, and Nteta 2019; Zhang, Sneyd, and Stevenson 2020; Kurpicz-Briki 2020; Chaloner and Maldonado 2019b).

For the execution of the WEAT on a word embedding, four sets of words are required: two equal sets of target words, X and Y, and two equal sets of attribute words, A and B. X and Y represent the words that are suspected to be biased, whereas A and B represent the attribute concept that the target words are suspected to be biased about (Mulsa and Spanakis 2020). The test statistic, $s(X, Y, A, B)$, measures the difference of the aggregated cosine similarities between the target word lists and the attribute word lists. It is formulated as:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

Where $s(w, A, B)$ measures the cosine similarities between the embedding of the target word, $\vec{w}$, with the embeddings of the attribute words, $\vec{a} \in A$ and $\vec{b} \in B$.

$$s(w, A, B) = mean_{a \in A} \cos(\vec{w}, \vec{a}) - mean_{b \in B} \cos(\vec{w}, \vec{b})$$

The significance (p-value) is computed by (1) merging the X and Y sets; (2) generating 100,000[5] or maximum possible distinct permutations ($PERM$) of this merged set ($X \cup Y$) by splitting each permutation ($i \in PERM$) into new pairs of X_i and Y_i sets, $(X_i, Y_i)_i$; and (3) performing a test statistic on each pair $(X_i, Y_i)_i$ and calculating the one-sided p-value of the permutation test:

$$p = \frac{\sum_{i \in PERM}[s(X_i, Y_i, A, B] > s(X, Y, A, B)]}{|PERM|}$$

Where p-value is the probability that the null hypothesis, H_0, is true. H_0 proposes that there is no difference between X and Y in terms of their relative cosine similarity to A and B. The high p-value ($p > 0.05$) suggests that the sets of target words, X and Y, are not significantly biased against the concept.

The test reports an effect size, d, based on the number of standard deviations that separate the two sets of target words according to their relationship with the attribute words (Caliskan, Bryson, and Narayanan 2017). The effect size is formulated as:

$$d = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{stddev_{w \in X \cup Y} s(w, A, B)}$$

For the integration of the WEAT experimental protocol in the Bias Assessment Module, the author utilised and extended the implementation provided by Chaloner and Maldonado (2019a).

3.2.2 Definition of Attribute and Target Words for the WEAT Experiments

Based on the findings and shortcomings summarised in Section 2.4, the author constructed 18 bias categories for testing with the Bias Assessment Module. For each of the categories, the sets of target words, X and Y, and the sets of attribute words, A and B, were constructed. In deciding which words to choose, the author was guided by:

1) Studies presented in Section 2.2.2 that measured gender, racial, religious, and age-related biases with the WEAT.
2) Studies in which implicit biases were measured using the IAT.
   Caliskan, Bryson, and Narayanan (2017) derived the target and attribute words for their study from research that measured prejudice using the IAT, the framework upon which the WEAT is built (as discussed in Section 2.2.1). In addition to examining the IAT from Project Implicit (2020), which originally developed the IAT, the author also examined several studies that measured gender and racial prejudice in children using the IAT. Thus, she was able to collect child-specific vocabulary for the WEATs.
3) Additional sources.
   Since there has only been one study that measured social stereotypes against animals with the WEAT (Caliskan, Bryson, and Narayanan 2017), and no psychological research that has evaluated stereotypes in children's literature about animals using the IAT, the author could not apply the first or second approach for the animal-related bias categories A2–A5. In this case, the author derived the target and attribute words from additional sources, such as studies on prejudice against animals in children's literature or the Thesaurus.

Table 2 provides an overview of the bias categories that were examined with the WEAT. For each category, the table specifies from where the author derived the target and attribute words. A table with the target and attribute words (Table 8) for each bias category can be found in the Appendix 9.1.

[5] 100,000 permutations were chosen in accordance with Chaloner and Maldonado (2019b) recommendation. This allowed the calculation of the test to remain computationally tractable.

Table 2: Overview of the categories to be tested with the Bias Assessment Module and the sources from which the attribute and target words were derived.

ID	Bias Category Name	Origin of the Attribute (AW) and Target (TW) Words
Regular Vocabulary		
Gender bias		
G1	career vs family	AW+TW: Chaloner and Maldonado (2019b), Caliskan, Bryson, and Narayanan (2017)
G2	maths vs arts	
G3	science vs arts	
G4	intelligence vs appearance	
G5	strength vs weakness	
Religious bias		
RL1	Christianity vs Islam	AW+TW: Project Implicit (2020)
RL2	Christianity vs Judaism	AW+TW: Project Implicit (2020)
RL3	Judaism vs Islam	AW+TW: Project Implicit (2020)
Age-related bias		
AG1	young vs old	AW: Thesaurus Hummert et al. (2002) + TW: Project Implicit (2020)
Animal-related stereotypes		
A1	flowers vs insects	AW + TW: Caliskan, Bryson, and Narayanan (2017)
A2	innocent sheep vs cruel wolf	AW: Wikipedia contributors (2020) TW: Wikipedia contributors (2020), Zito (2018)
A3	naïve bird vs clever fox	AW + TW Wikipedia contributors (2020), dictionary.com (2021)
A4	strong lion vs tender mouse	AW + TW Wikipedia contributors (2020), dictionary.com (2021)
A5	faithful dog vs selfish cat	AW+TW Wikipedia contributors (2020), dictionary.com (2021), Wehrmeyer (2010)
Child-Specific Vocabulary		
Racial bias		
CR1	European American vs African American	AW: Greenwald, McGhee, and Schwartz (1998); Caliskan, Bryson, and Narayanan (2017) + TW: IAT with children (Baron and Banaji 2006)
Gender bias		
CG1	math vs reading	AW: IAT with children Cvencek, Meltzoff, and Kapur (2014) + TW: IAT with children (Cvencek, Meltzoff, and Greenwald 2011)
CG2	math vs reading	AW: Caliskan, Bryson, and Narayanan (2017) + TW: IAT with children (Cvencek, Meltzoff, and Greenwald 2011)
Animal-related stereotypes		
CA1	flowers vs insects	AW+TW: Babcock et al. (2016) – IAT Test with children

3.2.3 Identifying New Sets of Target Words

Since the WEAT has limitations, which are described in Section 2.3, the author adapted the unsupervised clustering approach suggested by Chaloner and Maldonado (2019b)[6] to detect new target word sets from the vocabulary of the word embeddings. Algorithm 1 shows the procedure for identifying new sets of target words and assessing the results with the WEAT based on Chaloner and Maldonado (2019b) work. The author used this procedure as a basis for her implementations.

[6] Since Aran et al. (2020) and Swinger et al. (2019) do not provide implementation details for their approaches, the author decided to work with the method suggested by Chaloner and Maldonado (2019b) to detect new sets of biased words. The implementation of the method is publicly available on the GitHub platform (Chaloner and Maldonado 2019a).

Algorithm 1: Identifying New Sets of Target Words

1: $vocab$ vocabulary from the w2v model
2: $vectors$ word vectors from the w2v model
3: c clusters count
4: A, B sets of attribute words
5: n maximum number of words in a set of target words per cluster
6: $cluster\_labels$ ← create c clusters with the k-means++ algorithm from the $vectors$
7: $clusters$ ← group words from $vocab$ by $cluster\_labels$
8: initialise $scores$ to empty map
9: **for** each $cluster$ in $clusters$:
10: initialise X, Y as empty sets of target words
11: **for** each $word$ in $cluster$:
12: a_mean ← aggregate cosine similarity between $word$ to A
13: b_mean ← aggregate cosine similarity between $word$ to B
14: $score$ ← calculate the score as $b\_mean - a\_mean$
15: **if** $score < 0$ **then**
16: add $word$ to X
17: **else**
18: add $word$ to Y
19: **end if**
20: **end for**
21: In X, leave only n target words with the highest score
22: In Y, leave only n target words with the highest score
23: perform WEAT with X, Y, A, B
24: **end for**

3.3 Testing Procedure

The following section outlines the testing procedure the author followed to assess biases in children's books using the Bias Assessment Module.

Prior to performing the WEAT for each of the corpora presented in Section 3.1.1, the author used the Bias Assessment Module to run the WEAT on the GoogleNews word embeddings (Mikolov, Chen, et al. 2013) and the GAP (Webster et al. 2018) word embeddings in order to verify whether the implementation of the WEAT experimental protocol worked as expected and the results described by Chaloner and Maldonado (2019b), whose implementations were used as a basis for the Bias Assessment Module, are replicable.[7]

The author then pre-processed all the corpora presented in Section 3.1.1 using the procedure outlined in Section 3.1.2 and trained two sets of the word embeddings for each of the corpora using the Bias Assessment Module as described in Section 3.1.3. She then used the Bias Assessment Module to perform the WEAT for each of the corpora and for each of the bias categories with the target and attribute words defined in Section 3.2.2. The WEAT results were recorded to the respective files, which are described in Section 4.2.5.

In order to achieve representative results for the sub-corpora of the ChiLit and CLLIP corpora, the author created 10 unique ChiLit Small and CLLIP Small corpora using the Bias Assessment Module. For each of the sub-corpora, she trained a word embeddings model and performed the WEAT. Based on the WEAT results for all 10 sub-corpora, she then computed the average p-value and the average effect size across 10 results of each of the bias categories. The results that are presented in Section 5.2

[7] GoogleNews and GAP models are provided by Chaloner and Maldonado (2019b).

for the ChiLit Small and CLLIP Small are therefore the averaged results of the WEATs performed on the 10 sub-corpora.

For the CPB corpus, which is the corpus with the fewest words and – as the results of the WEATs presented in Section 5.2 reveal – the most out-of-vocabulary words, the author performed the unsupervised approach as explained in Section 4.2.4. Specifically, the Bias Assessment Module was used to create 100 clusters from the embedding of the CPB corpus. From each cluster, the target word set candidates were selected. The author then applied the WEAT on the CPB corpus with the child-specific attribute words from the categories CG1, CG2, and CA1, as well as the candidates for the target word sets found through unsupervised clustering. In order to keep the computation tractable, the author ran the WEATs with clusters for 1,000 permutations only.

3.4 Limitations

The following study had three limitations. First, the sample of children's literature only includes English books. This constraint was necessary since it would go beyond the scope of this study to include multiple languages of children's books, some of which might have been written in different cultural environments. Second, the corpora that were chosen for this study, especially the CPB, are relatively small, which may affect the quality of the word embeddings (Chakraborty, Badie, and Rudder 2016). However, there were no publicly available corpora consisting of English children's books that contain more words than the CLLIP or ChiLit corpora. Third, the ChiLit and CPB corpora have fluid boundaries. The latest book that the ChiLit corpus contains was published in the 20th century, and 30 of 100 books in CPB were published in the early 21st century. Since removing the books from the corpora would affect the quality of word embeddings, the author decided to leave all books in the corpora.

4 Bias Assessment Module

Based on the approaches discussed in Sections 2.2 and 2.3, as well as the experimental setup outlined in Section 3.2, this chapter provides an overview of the design and technical implementation of the Bias Assessment Module the author developed to measure biases in 19th and 20th century children's literature. The author presents the architecture of the module, followed by the description of the central workflows, such as module initialisation, corpus processing, model supplying, bias measuring with the WEAT, new category detection for biased target words and WEAT result logging.

The module is written in Python 3.7.4. The development was managed through the GitHub platform using the Git version control system. The code (Release v0.1) can be accessed on GitHub at the following address:

`github.com/vlebedynska/word-embeddings-childrens-books` (Olari 2020)

To improve the traceability of additions made over the course of this work, the author committed the progress of the work on a regular basis. A changelog is attached to Appendix 9.5.

4.1 System Architecture

The Bias Assessment Module is designed according to the object-oriented approach and has a modular structure that provides the possibility to extend the application with new corpora without changing the implementation. The following section provides an overview of the file and class structure of the module.

4.1.1 File Structure

The root directory of the project contains two main directories: `data` directory, holding single corpora and trained models, and `bias_assessment_module` directory, consisting of the source code. The root directory also stores a main `config.json` file. It contains configuration parameters for training the word embeddings, specifications for the unsupervised clustering using the k-means++ algorithm and the WEAT-lists with target and attribute words structured after bias categories for performing the WEATs.

In `data` directory, each corpus is stored in a separate directory and includes, besides the text or XML files containing the children's books, a `_config.json` file with the model-specific configuration parameters, such as a type of word embeddings model that it represents. There is also a `cache` directory where the trained models and the test results are saved.

4.1.2 Class Structure

Figure 1 demonstrates a simplified overview of the system components of the Bias Assessment Module and the relations between single components located in the `bias_assessment_module` directory. The following list provides a short description of each component:

- `BiasAssessmentModule` acts as a main interface for other components of the project and implements functionalities for starting the WEAT experimental protocol.
- `WeatConfig`, `ModelConfig` and `EmbeddingsClustererConfig` hold respective configuration data.
- `BiasAssessor` implements the WEAT experimental protocol.
- `TestResult` stores the WEAT results, including the p-value and Cohen's d.
- `BiasAssessorException` defines an exception in the case of an error occurring during the WEAT execution.
- `ModelHandler` encapsulates the mechanism which provides FastText or word2vec model instance.
- `CorpusSupplier` is an abstract class that defines functions for loading corpus data and implements the function for generating multiple sub-corpora from one corpus.
- `ModelSupplier` is an abstract class that defines functions for saving and loading the word embeddings.
- `ModelSupplierFactory` provides concrete implementations of the `ModelSupplier` class based on the model configuration.
- `ModelAndCorpusSupplier` implements `CorpusSupplier` and `ModelSupplier` abstract classes and provides concrete implementation for training the word2vec model. It saves and loads models using concrete implementations from the corresponding classes.
- `ChiLitSupplier`, `CLLIPSupplier`, `CPBSupplier`, `GoogleNewsSupplier` and `GAPSupplier` provide concrete implementations of the `ModelSupplier` class. `ChiLitSupplier`, `CLLIPSupplier` and `CPBSupplier` also implement the base class

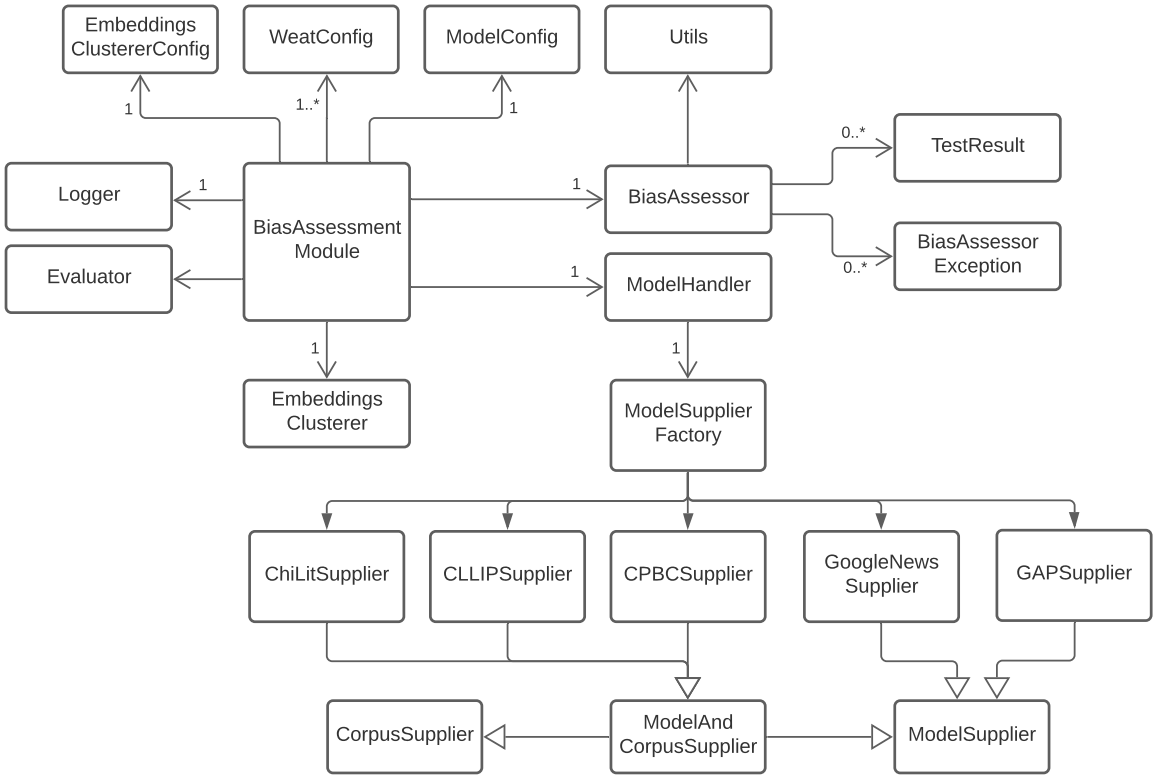

Figure 1: Class architecture of the Bias Assessment Module.

`ModelAndCorpusSupplier` and, thus, concrete functions for loading and saving the corresponding corpora, as these classes are also `CorpusSuppliers`. `GoogleNewsSupplier` and `GAPSupplier` do not implement `CorpusSupplier`, since the author does not have original corpus data and uses ready-trained models provided by Chaloner and Maldonado (2019b).

- `EmbeddingsClusterer` implements the unsupervised approach, as described in Section 3.2.3, using the k-means++ algorithm for clustering word embeddings. It also contains further functions to get the target words from the clusters and prepare the WEAT config.
- `Logger` provides logging functionality used for dumping test results into a file and on the console.
- `Evaluator` implements a helper function to calculate a mean score for multiple test results per bias category.
- `Utils` implements various help functions for the WEAT experimental protocol, for instance, to balance the sets of target and attribute words.

4.2 Implementation Details

The following section provides selected implementation details of the Bias Assessment Module. The implementation is additionally visualized as a sequence diagram in Figure 2 in Appendix 9.2. The diagram includes the central workflows and depicts the interaction between the module components in the example of the ChiLit corpus as the `ModelSupplier`. For sake of clarity, the presentation of loops and conditions is simplified, and exception handling is excluded.

4.2.1 Module Initialisation

The entry point into the Bias Assessment Module is the `main` method in the `WeatTester.py` file. It instantiates the object of the `BiasAssessmentModule` class with the parameters that the user has specified in the `config.json` and has additionally passed via the command line arguments on application start. The `config.json` file contains lists of pre-defined target and attribute words for WEATs, as described in Section 3.2.2, and default parameters that are used to train the word embeddings and perform clustering. The command line arguments are optional, and if passed, they override the default values in the `config.json` file. Table 3 shows the list of possible command line arguments.

Table 3: List of arguments that the user can input on start of the Bias Assessment Module.

Argument		**Description**
`-cr`	`--corpus`	Corpus name `CLLIP_Corpus`, `ChiLit_Corpus`, `CPB_Corpus`, `GAP`, `GoogleNews`
`-a`	`--amount`	Number of sub-corpora (applicable for `CLLIP_Corpus` and `ChiLit_Corpus` only, since they can be split into smaller sub-corpora)
`-s`	`--size`	Dimensionality of the word vectors*
`-p`	`--permutations`	Number of permutations for a single WEAT
`-w`	`--window`	Maximum distance between the current and predicted word within a sentence*
`-sg`	`--skipgram`	Training algorithm: `1` for skip-gram; `0` for CBOW*
`-m`	`--mode`	Running mode: `WEAT`, `Clustering`, `WEAT_and_Clustering`

* not applicable for GAP and GoogleNews

`WEATTester.py` merges the user input to the config data previously loaded from the `config.json` file. While instantiating, the `BiasAssessmentModule` processes the merge by extracting the data and storing it in objects with properties as shown in Table 4. This step is necessary to ensure the structured use of configuration data during the application's runtime.

Table 4: Configuration objects in `BiasAssessmentModule`.

Object name	**Type**	**Properties**
`model_config`	`ModelConfig`	`corpus path, model path, corpus_name,`

		amount_of_corpora, sg, epochs, number_of_permutations, size and window
model_clusterer_config	EmbeddingsClustererConfig	init, n_clusters, batch_size, max_no_improvement, verbose, cluster_words_count
[weat_configs]	WeatConfig	name, a, b, x, y

BiasAssessmentModule then creates single instances of the ModelHandler, BiasAssessor and Logger classes. Since the instantiation of the ModelHandler involves multiple steps, it is explained in detail in Section 4.2.2. The instance of the BiasAssessor is created with a list of models provided by the ModelHandler. Usually, the list consists of only one model trained on the entire corpus. However, if the user has specified --amount argument at the start of the application or the amount_of_corpora parameter in the config.json, then the list contains the specified number of models trained on the sub-corpora of the corpus. The instance of the Logger class is created with the model_id, which is also provided by the ModelHandler. To this end, the main method owns a module – an initialised instance of the BiasAssessmentModule class.

4.2.2 Corpus Processing and Supplying the Models

Supplying the model involves multiple steps, which differ depending on the model_type of the chosen corpus. On the instantiating of the ModelHandler, the internal _load() function is called. It creates a ModelSupplier instance via the helper class ModelSupplierFactory which is able to instantiate all possible model types: ChiLit, CLLIP, CPB, GoogleNews and GAP. ModelSupplier defines methods that encapsulate saving and loading the models which are then implemented by concrete model suppliers: ChiLitSupplier, CLLIPSupplier, CPBSupplier, GoogleNewsSupplier and GAPSupplier.

In the exemplarily case of the ChiLit corpus, the ModelSupplierFactory instantiates the ChiLitSupplier object and returns its instance to the ModelHandler. The ModelHandler then checks, via the model_id provided by the ModelAndCorpusSupplier, if the model for the ChiLit corpus already exists in the cache. To make the model_id unique, the model_id is built from the parameters specified in the model_config object.

If there is a directory in cache named after the model_id, all models inside the directory are loaded and returned to the ModelHandler. If there is no corresponding directory, the new models are created. For this, the corpus is pre-processed and submitted for training with the parameters specified in the model_config object using the word2vec algorithm. In the case of the ChiLit corpus, the corpus is pre-processed by removing the author and title lines from each book and tokenising the text using the simple_preprocess() function from the gensim library (Řehůřek 2021). If the –amount_of_corpora configuration parameter has the value greater than one, the sub-corpora are created from the corpus, and for each sub-corpus, a word2vec model is trained. After the training is completed, the models are saved to cache in a new directory named after the model_id and returned to the ModelHandler. The initialisation of the ModelHandler is finished.

4.2.3 Execution of the WEAT Experimental Protocol

If the --mode parameter was specified as "WEAT" or "WEAT_and_Clustering", the WEAT experimental protocol is started by calling the run_weat() function after the initialisation of the BiasAssessmentModule. As a parameter, the function gets a list of strings which must be the valid names of the WEAT categories loaded from the config.json into the BiasAssessmentModule at the initialisation step described in Section 4.2.1. For each name, the run_weat() function gets corresponding weat_config objects. After this mapping, the _run_weat_intern() method is called with the list of weat_config objects as a parameter. It creates the logging files[8] and starts the run_bias_test() function for each weat_config object in BiasAssessor. The run_bias_test() performs WEAT hypothesis protocol for all models.

[8] If the files already exist, for instance, from the previous test run, the data will be removed.

The implementation of the WEAT experimental protocol is built upon the implementation proposed by Chaloner and Maldonado (2019b). The following presents the main steps that the `BiasAssessor` performs for each model, while the algorithm itself is presented in detail in Section 3.2.1.

1) The words from the target and attribute word lists stored in the `weat_config` under `x`, `y` and `a`, `b` that are not in the corpus vocabulary are filtered out. If one of the lists is empty after the filtering process, an error is raised, and the WEAT is finished.
2) If sets of target words `x` and `y` or of attribute words `a` and `b` become unequal after the filtering procedure, they are balanced by reducing the number of words in the category with the larger number of words to the smaller number of words in the other category. Although this step was omitted by Chaloner and Maldonado (2019b), the author introduced it in her implementation, since it was initially intended by Caliskan, Bryson, and Narayanan (2017).
3) The p-value is computed as described in Section 3.2.1 for a number of permutations specified in `model_config`.
4) Effect size (Cohen's d) is computed as outlined in Section 3.2.1.
5) `TestResult` object is created with the results of the WEAT.

At the end, `TestResults` calculated for each model are aggregated and returned to the `BiasAssessmentModule` for evaluation and logging.

4.2.4 Finding New Target Words for Specific Bias Categories with Clustering

If the user has selected `Clustering` or `WEAT_and_Clustering` in the `--mode` argument upon starting the Bias Assessment Module, the algorithm for identifying new sets of biased target words is performed as described in Section 3.2.3. It is started after the initialisation of the `BiasAssessmentModule` by calling the `run_weat_with_clusters()` function. As a parameter, the function gets a list of strings which must be the valid names of the WEAT categories loaded from the `config.json` into the `BiasAssessmentModule` at the initialisation step described in Section 4.2.1. In the first step, the function instantiates the `EmbeddingsClusterer` object by applying the `MiniBatchKMeans` algorithm to the model vectors and assigning each word from the model vocabulary to a cluster.

In the second step, for each WEAT category in the list passed to `run_weat_with_clusters()`, `x` and `y` target words are found from the clusters, as described in Section 3.2.3. In the third step, these `x` and `y` word sets are used as target words for the WEAT in combination with the `a` and `b` attribute words provided by the `weat_config` of the current category.

4.2.5 Logging the Test Results

Before dumping the WEAT results into a file, the results for each WEAT and corpus are averaged, and the mean test result is calculated. The `Logger` class that was created upon the module initialisation dumps the results of the WEAT, as shown in Table 5. If an error raises during the execution of the WEAT, it is caught and logged to the `.log` file.

Table 5: File names and contents for the storing of the WEAT results.

File Names	File Contents
`_results.txt`	`model_id`, `bias_category`, averaged `p_value`, averaged `cohens_d`, `number_of_permutations`, averaged `total_time`
`_full_results.txt`	`model_id`, `bias_category`, `p_value`, `cohens_d`, `number_of_permutations`, `total_time`, `absent_words` (absent attribute and target words), `used_words` (used attribute and target words)

5 Results

The following chapter provides insights into the results of the study and shows discrepancies in the significance of prejudice when comparing sets of word embeddings trained on 19th and 20th century children's books. The chapter is structured after the testing procedure defined in Section 3.3 that the author followed while conducting the experiments with the Bias Assessment Module. First, the results of the WEAT for the GoogleNews and GAP embeddings are compared with the findings from previous

studies. Second, the results of the WEATs for each bias category and for each corpus are presented. Third, the author presents the target words found through the unsupervised approach applied to the CPB corpus.

5.1 Replicating Results from Previous Studies

To verify the correctness of the implementation of the WEAT algorithm adapted from Chaloner and Maldonado (2019b), the author performed the G1–G5 WEATs on GoogleNews and GAP corpora. Table 6 shows the results obtained in the current study in comparison to the results provided by Chaloner and Maldonado (2019b).

For the GoogleNews corpus, all WEAT results provided by Chaloner and Maldonado (2019b) could be replicated. For the GAP corpus, the author could also replicate the results for the G1, G3 and G4 categories. The G2 category demonstrated a strong bias that was not present in the original findings, and the G5 category did not exhibit bias, as suggested by Chaloner and Maldonado (2019b).

Table 6: WEAT results for GoogleNews and GAP corpus for G1–G5 categories obtained in the current study (on the left) in comparison to the original findings (on the right) published by Chaloner and Maldonado (2019b). The p-values in bold indicate statistically significant bias, $p < 0.05$.

	GoogleNews		GAP				GoogleNews		GAP	
	p	d	p	d			p	d	p	d
G1	**0.001**	1.37	**0.007**	1.19		B1: career vs family	**0.0012**	1.37	**0.0015**	1.44
G2	**0.017**	1.02	**0.000**	1.09		B2: math vs arts	**0.0173**	1.02	0.0957	1.04
G3	**0.004**	1.25	0.116	0.79		B3: science vs arts	**0.0044**	1.25	0.1434	0.71
G4	**0.000**	0.98	0.550	-0.97		B4: intelligence vs appearance	**0.0001**	0.98	0.9988	-0.64
G5	**0.006**	0.89	0.141	0.64		B5: strength vs weakness	**0.0059**	0.89	**0.0018**	0.77

5.2 WEAT Results for 19th and 20th Century's Children's Literature

Overall, the author conducted WEATs for all bias categories outlined in Section 3.2.2 on ChiLit and ChiLit Small corpora representing 19th century children's literature and on CLLIP, CLLIP Small and CPB corpora representing 20th century children's literature. In the following sub-sections, the author describes the results shown in Table 9 in Appendix 9.3 and Table 10 in Appendix 9.4.

Table 9 provides an overview of all WEATs results, in which p-values in bold indicate statistically significant bias, $p < 0.05$. The cells with '-' indicate that the WEAT test could not be performed for this corpus and category because one of the target or attribute word sets was missing. Table 10 shows the number of out-of-vocabulary target and attribute words in ChiLit, CLLIP and CPB embeddings. The tests that were not performed due to lack of vocabulary are filled in with black. Since ChiLit Small and CLLIP Small were randomly generated 10 times and each time contained a different amount of out-of-vocabulary target and attribute words, they are not included in Table 10.

5.2.1 Gender Bias

The ChiLit corpus was found to exhibit significant results for the G1 and G4 categories when trained using the context window size of 30 words. This result indicates that women in 19th century children's literature are more associated with family and appearance, while men are more related to career and intelligence. For the CLLIP corpus, the presence of bias in the G4 category could also be proven. Additionally, the findings indicate that CLLIP embeddings, when trained with the context window size of 10 words, exhibit bias in the G5 category. Men are more associated with strength in 20th century children's books, whereas women are represented as weak and fragile. Compared to ChiLit and CLLIP corpora, CPB corpus did not show any bias for the categories G1–G5.

The WEAT results for the CG1 category confirmed the presence of bias in four of five corpora. The names of boys are significantly more often associated with numbers and math than the names of girls, and the names of girls are more related to books, reading and stories than the names of boys. For the ChiLit, ChiLit Small and CLLIP corpora, this result could be confirmed regardless of the context window.

The CPB corpus also demonstrated a significant p-value in both context windows for the CG2 category, with the effect size varying from strongly negative for windows size 10 to strongly positive for window size 30. This means that, in a context window of 10, boys are more associated with books,

stories and letters and girls with numbers, and in the context window of 30, vice versa – the girls are more associated with books and stories, while boys are more related to numbers.

Compared to the ChiLit and CLLIP corpora, the CPB had the most out-of-vocabulary words in gender related tests. For instance, Table 10 reveals that it contained only three of four words in the X target word set for the CG2 test. Several WEATs could not be performed with the ChiLit Small and CLLIP Small, since the target or attribute word sets were missing in the vocabulary of the sub-corpora. The WEATs that were performed with the ChiLit Small and CLLIP Small did not show any significant results.

5.2.2 Religious Bias

While running WEATs for categories RL1–RL3, it was found that CPB, ChiLit Small and CLLIP Small lacked target word sets for Islam and Judaism concepts. Therefore, the WEATs tests could only be performed with the ChiLit and CLLIP corpora. In both corpora, more than half of the words in the target word sets were missing for the RL3 category.

WEAT results for RL1 bias categories indicate that there is a significant religious bias in 19th and 20th century children's books; both ChiLit and CLLIP corpus showed that Christianity is more likely than Islam to be closer to pleasant than unpleasant attributes. However, the religious bias was no longer detectable in these corpora when the embeddings were trained with a higher context window size.

5.2.3 Animal Stereotypes

Stereotypes towards animals cannot be proven for 19th century children's literature, since all WEATs returned unsignificant results. In 20th century children's books, three results are significant.

The CLLIP corpus revealed that birds are often described as naïve and shy, while foxes are represented as intelligent and cruel. This result only applied to the CLLIP embeddings trained with the context size window of 30 words. The CPB corpus exhibits biases in the CA1 category. However, similar to the results described for the gender bias in Section 5.2.1, these findings are contradictory. The effect size is strongly positive for the CBP embeddings trained with the context size window of 10 words and is strongly negative for the CBPC embeddings trained with the context size window of 30 words.

Similar to findings in Sections 5.2.1, 5.2.2 and 5.2.3, ChiLit Small and CLLIP Small miss several words used in target word sets in the A1–A5 and CA1 categories. The tests that were performed did not show any significant results.

5.2.4 Age-related and Racial Biases

The result for WEAT in the AG1 category indicates that the language used in 19th and 20th century children's books does not exhibit age-related bias. Young people and old people are equally close to pleasant and unpleasant concepts. Due to the lacking vocabulary, the AG1 test could not be performed on the ChiLit Small or CLLIP Small corpora.

The WEAT for the CR1 category showed that both the ChiLit corpus with 19th century children's books and the CLLIP corpus with 20th century children's books incorporate racial biases. European American names have a stronger association with pleasant and positive concepts than African American names. African American names are more likely than European American to be associated with negative and unpleasant attributes than with pleasant. This result was also demonstrated for the ChiLit Small corpus.

Table 10 indicates that CLLIP and ChiLit corpus contained three of four words missing in the attribute word set B. The CLLIP Small and CPB corpora did not have sufficient vocabulary in target word sets to conduct the WEATs.

5.3 Results for Detecting New Bias Word Categories in CPB Corpus

Since the CPB corpus had the most words out-of-vocabulary compared to the ChiLit and CLLIP corpora, the unsupervised method outlined in Section 3.2.3 was applied to categories CG1, CG2 and CA1, which build upon the child-specific attribute words. The CR1 category could not be tested, since one of the attribute word categories was completely missing. Table 7 shows some of the findings, which returned significant results $p < 0.001$.

Table 7: Selection of induced CG1, CG2 and CA1 biased word categories per cluster (an extract).

CG1	**Male Names**	**Female Names**	***d***
	steam, trout, discovered, squeezed, flipped, arrived, muddy, backward	silently, terribly, anybody, fire, scarf, grandpa, strangest, streets	1.93
	games, race, princess, chopping	guess, feeling, frighten, last	1.54
	sticks, smelling, siren, troubles, toppings, cough, roller, bumble, sending	carrying, yet, evening, blanket, comes, disappear, clear, salad, stellaluna	1.53
	john, tom, comfortable, hug, africa, towards, shot, mulligan, delight, reporters	alice, pretty, knock, okay, drop, splashing, front, leaned, basket, fascinated	1.87
CG2	**Male attributes**	**Female Attributes**	
	excitement, boy, school, fuzzy, joined	gentle, kisses, ambulance, elbow, games	1.89
	remarkable, jumping	dining, sting	1.73
	concluded, sleepy, chirped, showed, isn, jumped, snowball, stepped, fruit, pumped	listen, ridden, driver, moan, conversation, imagine, easily, remind, sam, blah	1.94
CA1	**Flowers**	**Insects**	
	race, weak, poor, ambulance, aside, gentle, townspeople	scare, chopping, elbow, games, few, feeling, kisses	1.52
	hadrosaur, celery, tunnels, fixed, waits, fair, mischief, dug, mothers, pickle	excellent, wakes, suggested, queen, serious, huggy, shooter, monster, wow, grace	1.76
	stopping, shoe, pillow, spots, brains, wound, changed, ahhh, clocks, congratulations	plains, zebra, speck, striped, winning, floors, chug, coconut, drawing, peppers	1.70

6 Discussion

Considering the research questions presented in Section 1.1 and the shortcomings summarised in Section 2.4, in this chapter, the author discusses the results presented in Chapter 5. To begin, the author summarises the main findings. The author then compares the findings for the 19th century with the findings for the 20th century and reflects on them using previous research that has examined prejudice in 19th and 20th century children's literature. Finally, the author evaluates the study's outcomes with regard to the chosen approaches and discusses the possible factors that may have influenced the results.

6.1 Summary of and Reflections on the Results

In summary, the results presented in Chapter 5 indicate that the implementation of the WEAT algorithm is correct and the embeddings trained on children's literature exhibit gender, religious and racial biases. These findings partially correspond with the previous research on children's literature presented in Section 2.1:

1) Similar to the conclusions of Anderson (2013) and Turner-Bowker (1996), the results outlined in Section 5.2.1 suggest that boys are significantly more often associated with math and science than girls. The girls are also more related to weakness than boys, and the boys are closer to the attribute of strength than girls.
2) The evidence that African American names are more commonly associated with unpleasant concepts while European American names are linked with pleasant concepts, as presented in Section 5.2.4, supports the findings provided by MacCann (2013) and proves that racial bias is apparent in children's literature.
3) Contrary to the assertions of several researchers who stated that religion has almost disappeared from children's literature (Ghesquière 2005; Hunt 2005; Trousdale 2011), the statistical results outlined in Section 5.2.4 demonstrated that religion still exists in books read by children. The findings indicated that Christianity is much more likely than Islam to be closer associated with pleasant concepts than to unpleasant.
4) The presence of age-related biases could not be confirmed. This finding contrasts with previous research that claimed that older people are portrayed at a disadvantage in children's books compared to other adults (McGuire 2016; Taylor 1980; Crawford 2000).

5) The presence of stereotypes against animals in children's literature, as supposed by Dunn (2011) and Wehrmeyer (2010), could not be confirmed, with the exception of the stereotype that birds are preferably described as naïve and shy, while foxes are portrayed as intelligent and cruel.

When comparing the WEATs results on the ChiLit embeddings with the results on the CLLIP and CPB embeddings, it was observed that the gender bias is consistent throughout the 19th and 20th centuries. Although some researchers have observed a decrease in gender bias since 1800 (Turner-Bowker 1996; Jones et al. 2020), the WEAT results outlined in Section 5.2.1 show that gender bias is still present, and even stronger, in the 20th century. While only one of seven gender categories was biased in the ChiLit embeddings, four were biased in the CLLIP and CPB embeddings.

New target word sets presented in Section 5.3 returned significant WEATs results for the gender bias CG1 and CG2 categories when tested on the CPB embeddings. However, the target words found seem to be detached, and it is difficult to summarize which concepts from the 20th century they represent. The word selection in each of the sets seems to be random and incohesive.

Regarding the religious and racial biases in language, there is no discrepancy in the results between the 19th and 20th centuries. Despite the increasing secularisation (Hunt 2005; Trousdale 2011) and diversity in society (Long 1984), Christianity is presented more positively than Islam throughout the centuries, and European American names are preferred over African American names.

Biases inherent in the portrayal of animals were not detected in 19th century children's literature. However, they seem to be present in children's books published in the 20th century. Thus, the stereotypes measured in the CPB corpus are contradictory due to the strong fluctuation of the effect size.

Similar to the results presented for gender bias, the new target words found for the insects-flower category through unsupervised clustering returned significant results for the CA1 category when tested on the CPB corpus. However, it is difficult to summarize what stereotypical concept of insects or flowers these target words represent.

Several WEATs returned unsignificant results. For instance, 19th and 20th century children's books did not exhibit gender bias regarding the career–family (G1) and study subjects (G2 and G3) bias categories, although these biases were discovered and confirmed several times in previous research on corpora from various domains (Chaloner and Maldonado 2019b; Caliskan, Bryson, and Narayanan 2017). The preference of young over old people tested in the AG1 bias category could not be confirmed either for the 19th or 20th centuries. In comparison to the ChiLit and CLLIP corpora, the ChiLit Small and CLLIP Small yielded unsignificant results in almost all bias categories. This fact again emphasizes the importance for the size of the corpus for using it with the WEAT.

6.2 Evaluation of the Study Results

It is possible that the biases in categories G1–G3, RL2–RL3, A1, A2, A4, A5, AG1 are not present in children's books. However, there are other reasons for why the biases could not be captured, which also raise concerns about the overall reliability of the study results.

Section 5.2 has shown that target and attribute words were partially lacking in almost all WEATs. According Chaloner and Maldonado (2019b), this difficulty could have influenced the low bias measurement. In several WEATs run on the ChiLit Small, CLLIP Small and CPB corpora, the target and attribute words were completely missing, so WEATs could not be performed at all. At this point, the assertion that the size of the corpus is crucial for the efficacy of word embeddings is again confirmed (Chakraborty, Badie, and Rudder 2016).

Most target and attribute word pairs used for the tests were not child specific. They were derived from previous studies measuring biases with the WEAT. However, the language used in children's books differs from the language used in literature for adults. An indication that the selected words might have misrepresented the concept is provided through comparison of the results of the G3 and CG1 tests. To measure the gender bias in the G3 category, the attribute and target words provided by Chaloner and Maldonado (2019b) were used. The results were not significant for any of the children's books embeddings. However, gender bias measured with child-specific vocabulary derived from the IATs with children revealed the presence of the bias in three of five corpora. Therefore, the words used for the bias measurement in the G3 category might have failed to represent the biased concept.

Another issue that raises concerns about the study results for the WEATs in the A2–A5 category is whether the target and attribute words chosen by the author accurately reflect the alleged stereotypes.

Indeed, the representation of animals is often ambiguous. For instance, the author decided to describe the fox as a dangerous and evil animal. However, foxes are also portrayed as good and noble character in fables and legends (Wikipedia contributors 2020). Because there are few studies examining biases against animals in children's books, more research is needed in the future to more precisely delineate these biased categories.

The fact that the set of target or attribute words with the larger vocabulary is randomly aligned with the set with smaller vocabulary at the beginning of the WEAT leads to unstable results for the WEAT. Each WEAT run produces different result output. Since it is unclear what amount of target and attribute words is acceptable to achieve decent results with the WEAT, the impact of having unequal and limited sizes of target or attribute word sets is difficult to estimate. Caliskan, Bryson, and Narayanan (2017) used sets consisting of 8 to 25 words. However, it is unclear whether the sets containing less words are sufficient for conclusive results. If that is not the case, the categories with originally smaller vocabulary are probably unsuitable candidates for a WEAT.

Section 5.2 revealed that biases appeared and disappeared in the same corpus depending on the context size window. However, sometimes they remained stable in both variants of corpus embeddings. Although it is known that the smaller context window size captures more information about singular words and larger the domain information (Levy and Goldberg 2014), it still questionable whether the results that were confirmed for both context size windows are more conclusive than those that were confirmed for one and not the other.

7 Conclusion

This study aimed to examine biases in large amounts of 19th and 20th century children's literature by means of word embeddings and the WEAT. In her research proposal, the author first intended to explore the extent to which known implicit biases can be found in word embeddings trained on corpora consisting of children's books. Second, she aimed to identify new categories with biased words and, third, to investigate the discrepancies in the significance of prejudice when comparing sets of word embeddings trained on 19th and 20th century children's literature.

Since none of the previous studies examined biases in word embeddings trained on children's literature with the WEAT, the author studied the previous research on biases in children's literature and investigated the feasibility of measuring these biases in word embeddings with the WEAT. She also investigated the limitations of the WEAT when applied to word embeddings trained on domain-specific corpora.

Having established the theoretical framework, the author developed a Bias Assessment Module to answer the posed research questions. She used it to process the children's books corpora, train the word embeddings, detect new biased target words and measure biases with the WEAT.

The answer to the first research question is that biases known from the previous research could be found to some extent. While findings on gender and racial biases could be replicated in word embeddings trained on children's books, the age-related bias and stereotypes on animals could largely not be confirmed. In contrast to the findings from previous research, the results indicated the presence of a religious bias.

To answer the second research question, the author applied an unsupervised clustering on the corpus with the lowest vocabulary. In Section 5.3, an excerpt containing the newly found target word sets was presented. Although the target word sets found returned significant results when used for the WEAT with the corpus they originated from, the selection of words in the sets did not seem to be cohesive.

Although the third research question was comprehensively answered in Section 6.1, the discrepancies in the significance of prejudice are briefly summarised here. When comparing WEAT results for the word embeddings trained on the ChiLit corpus with those trained on the CLLIP corpus, the discrepancies are minor. Compared to the ChiLit embedding, the CLLIP embedding showed bias in two additional gender categories and one animal-related category. Other results appear to be steady for both the ChiLit and CLLIP embeddings. Racial, religious, and gender biases related to study subjects are consistently significant, whereas the age-related bias is not. However, when the results for the ChiLit embedding are compared to the results of the CPB embedding, the discrepancies are more prominent.

Only the result of the gender category related to the study subjects is significantly high for both of the embeddings.

The results for the CPB embedding, as well as the ChiLit Small and CLLIP Small embeddings, show that a sufficiently large vocabulary of a corpus is essential in performing the WEAT. Therefore, future research should investigate biases in larger corpora of children's books. It can also explore a child-specific delineation of words used for the WEATs and further techniques than the WEAT to assess bias. Since the results for the new target word sets were not cohesive, future research can also focus on testing further unsupervised methods for finding new biased word sets on the children's books embeddings.

8 Bibliography

Anderson, C. G. 2013. "Do Quality Informational Children's Books Show a Gender Bias? A Pictorial Examination of ten Years of Sibert and Orbis Pictus award Winners." In.

Aran, X. F., T. V. Nuenen, N. Criado, and J. M. Such. 2020. 'Discovering and Interpreting Conceptual Biases in Online Communities', *ArXiv*, abs/2010.14448.

Azmiry, N. 2014. 'Animals and Their Functions in Children's Literature Since 1900', University of Liberal Arts Bangladesh.

Babaeianjelodar, M., S. Lorenz, J. Gordon, J. Matthews, and E. Freitag. 2020. "Quantifying Gender Bias in Different Corpora." In *Companion Proceedings of the Web Conference 2020*, 752–59. Taipei, Taiwan: Association for Computing Machinery. 10.1145/3366424.3383559.

Babcock, R. L., E. E. MaloneBeach, J. Hannighofer, and B. Woodworth-Hou. 2016. 'Development of a Children's IAT to Measure Bias Against the Elderly', *Journal of Intergenerational Relationships*, 14: 167-78. 10.1080/15350770.2016.1195245.

Barman, M., A. Awekar, and S. Kothari. 2019. 'Decoding The Style And Bias of Song Lyrics', *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Baron, A. S., and M. R. Banaji. 2006. 'The Development of Implicit Attitudes:Evidence of Race Evaluations From Ages 6 and 10 and Adulthood', *Psychological Science*, 17: 53-58. 10.1111/j.1467-9280.2005.01664.x.

BNC Consortium. 2007. 'British National Corpus, XML edition', Accessed 28.12.2020. https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554.

Burke, C. L., and J. G. Copenhaver. 2004. 'Animals as people in children's literature', *Language Arts*, 81: 205-13.

Caliskan, A., J. Bryson, and A. Narayanan. 2016. 'Semantics derived automatically from language corpora necessarily contain human biases', *Science*, 356. 10.1126/science.aal4230.

———. 2017. 'Semantics derived automatically from language corpora contain human-like biases', *Science*, 356: 183-86.

Centre for Corpus Research. 2019. 'ChiLit Corpus on GitHub', Accessed 27.12.2020. https://github.com/birmingham-ccr/corpora/tree/master/ChiLit.

Čermáková, A. 2017. 'The GLARE 19th Century Children's Literature Corpus in CLiC [Blog post]'.

———. 2018. "ChiLit: the GLARE 19th Century Children's Literature corpus in CLiC." In *GLARE project blog*. University of Birmingam.

Chakraborty, T., G. Badie, and B. Rudder. 2016. "Reducing gender bias in word embeddings." In.

Chaloner, K., and A. Maldonado. 2019a. 'GitHub Repository alfredomg/GeBNLP2019', Accessed 12.12.2020. https://github.com/alfredomg/GeBNLP2019.

———. 2019b. "Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories." In, 25-32. 10.18653/v1/W19-3804.

Child, I. L., E. H. Potter, and E. M. Levine. 1946. 'Children's textbooks and personality development: An exploration in the social psychology of education', *Psychological Monographs*, 60: i-54. 10.1037/h0093569.

Cole, E. M., and D. P. Valentine. 2000. 'Multiethnic Children Portrayed in Children's Picture Books', *Child and Adolescent Social Work Journal*, 17: 305-17. 10.1023/A:1007550124043.

Crawford, P. A. 2000. 'Crossing boundaries: Addressing ageism through children's books', *Reading Horizons: A Journal of Literacy and Language Arts*, 40: 1.

Cvencek, D., A. N. Meltzoff, and A. G. Greenwald. 2011. 'Math–Gender Stereotypes in Elementary School Children', *Child Development*, 82: 766-79. 10.1111/j.1467-8624.2010.01529.x.

Cvencek, D., A. N. Meltzoff, and M. Kapur. 2014. 'Cognitive consistency and math–gender stereotypes in Singaporean children', *Journal of Experimental Child Psychology*, 117: 73-91. https://doi.org/10.1016/j.jecp.2013.07.018.

Darni, F. I. N. A. 2017. "Gender Bias in Elementary School Language Textbooks." In.

Derman-Sparks, L., C. M. Goins, and J. O. Edwards. 2020. *Anti-Bias Education for Young Children and Ourselves* (National Association for the Education of Young Children).
dictionary.com. 2021. 'thesaurus.com'. https://www.thesaurus.com/.
Dunn, E. A. 2011. 'Talking animals: A literature review of anthropomorphism in children's books'.
Eberhardt, V. M. 2018. "Representations of Religion and Culture in Childreé s Literature." In.
Ethayarajh, K., D. Duvenaud, and G. Hirst. 2019. 'Understanding Undesirable Word Embedding Associations', *ArXiv*, abs/1908.06361.
Ghesquière, R. 2005. 'Hidden Religious Themes in 20th-Century European', *Religion, Children's Literature, and Modernity in Western Europe, 1750-2000*, 3: 305.
Google Code Archive. 2013. 'word2vec', Accessed 31.12.2020. https://code.google.com/archive/p/word2vec/.
Greenwald, A. G., D. E. McGhee, and J. L. Schwartz. 1998. 'Measuring individual differences in implicit cognition: the implicit association test', *Journal of Personality and Social Psychology*, 74: 1464.
Heywood, S. 2020. "Gender in children's literature in Europe." In *Encyclopédie pour une histoire numérique de l'Europe [online]*.
Hummert, M. L., T. A. Garstka, L. T. O'Brien, A. G. Greenwald, and D. S. Mellott. 2002. 'Using the implicit association test to measure age differences in implicit social cognitions', *Psychol Aging*, 17: 482-95. 10.1037//0882-7974.17.3.482.
Hunt, P. 2005. 'The Loss of the Father and sof God in English-Language Children's Literature (1800–2000)', *Religion, Children's Literature, and Modernity in Western Europe, 1750-2000*, 3: 295.
Jones, J. J., M. Amin, J. Kim, and S. Skiena. 2020. 'Stereotypical Gender Associations in Language Have Decreased Over Time', *Sociological Science*, 7: 1-35.
Kurita, K., N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. 2019. 'Measuring bias in contextualized word representations', *arXiv preprint arXiv:1906.07337*.
Kurpicz-Briki, M. 2020. 'Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings', *Arbor-ciencia Pensamiento Y Cultura*.
Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald. 2007. 'Understanding and using the implicit association test: IV', *Implicit measures of attitudes*: 59-102.
Leahy, M., and B. Foley. 2018. 'Diversity in Children's Literature', *World Journal of Educational Research*, 5: 172. 10.22158/wjer.v5n2p172.
Levy, O., and Y. Goldberg. 2014. "Dependency-Based Word Embeddings." In *ACL*.
Long, M. A. 1984. 'The Interracial Family in Children's Literature', *Interracial Books for Children Bulletin*, 15: 13-15.
MacCann, D. 2013. *White supremacy in children's literature: Characterizations of African Americans, 1830-1900* (Routledge).
McArthur, L. Z., and S. V. Eisen. 1976. 'Achievements of male and female storybook characters as determinants of achievement behavior by boys and girls', *Journal of Personality and Social Psychology*, 33: 467-73. 10.1037/0022-3514.33.4.467.
McCabe, J., E. Fairchild, L. Grauerholz, B. Pescosolido, and D. Tope. 2011. 'Gender in Twentieth-Century Children's Books', *Gender & Society - GENDER SOC*, 25: 197-226. 10.1177/0891243211398358.
McGuire, S. L. 2016. 'Early Children's Literature and Aging', *Creative Education*.
Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*.
Mikolov, T., I. Sutskever, K. Chen, G. s. Corrado, and J. Dean. 2013. 'Distributed Representations of Words and Phrases and their Compositionality', *Advances in Neural Information Processing Systems*, 26.
Montag, J., M. Jones, and L. Smith. 2015. 'The Words Children Hear: Picture Books and the Statistics for Language Learning', *Psychological Science*, 26. 10.1177/0956797615594361.
Morgan, H., and D. E. Forest. 2016. 'What Educators Need to Do with Biased Childreé s Books on Religion, Gender and Race', *Journal of International Social Studies*, 6: 74-83.
Mulsa, R. A. C. a., and G. Spanakis. 2020. 'Evaluating Bias In Dutch Word Embeddings', *ArXiv*, abs/2011.00244.

Olari, V. 2020. 'Bias Assessment Module', Accessed 11.01.2021. https://github.com/vlebedynska/word-embeddings-childrens-books.
Popovic, R., F. Lemmerich, and M. Strohmaier. 2020. "Joint Multiclass Debiasing of Word Embeddings." In *ISMIS*.
PrÈcenth, R. 2019. "Word Embeddings and Gender Stereotypes in Swedish and English." In.
Project Implicit. 2020. 'Project Impicit – Take a Test – Harvard Implicit Association Test (United States)', Harvard, Accessed 05.01.2021. https://implicit.harvard.edu/implicit/takeatest.html.
Raabe, T., and A. Beelmann. 2011. 'Development of Ethnic, Racial, and National Prejudice in Childhood and Adolescence: A Multinational Meta-Analysis of Age Differences', *Child Development*, 82: 1715-37. 10.1111/j.1467-8624.2011.01668.x.
Řehůřek, R. 2021. 'Gensim Library', Accessed 25.01.2021. https://radimrehurek.com/gensim/.
Rice, D., J. Rhodes, and T. M. Nteta. 2019. 'Racial bias in legal language', *Research & Politics*, 6.
Sketch Engine. 2020. 'British National Corpus', Accessed 28.12.2020. https://www.sketchengine.eu/british-national-corpus/.
Swinger, N., M. De-Arteaga, I. NeilThomasHeffernan, M. D. M. Leiserson, and A. Kalai. 2019. 'What are the Biases in My Word Embedding?', *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.
Tan, Y. C., and L. E. Celis. 2019. "Assessing social and intersectional biases in contextualized word representations." In *Advances in Neural Information Processing Systems*, 13230-41.
Taylor, G. C. 1980. 'ERIC/RCS: Images of the Elderly in Children's Literature', *The Reading Teacher*, 34: 344-47.
The Editors of Encyclopaedia Britannica. 2019. 'Abolitionism', Encyclopædia Britannica Accessed 20.01.2021. https://www.britannica.com/topic/abolitionism-European-and-American-social-movement.
Thompson, P., and A. Sealey. 2007. 'Through children's eyes?: Corpus evidence of the features of children's literature', *International Journal of Corpus Linguistics*, 12: 1-23. 10.1075/ijcl.12.1.03tho.
Trousdale, A. M. 2011. 'Honouring the questions: shifts in the treatment of religion in childreé s literature', *International Journal of Children's Spirituality*, 16: 219 - 32.
Turner-Bowker, D. M. 1996. 'Gender stereotyped descriptors in children's picture books: Does “curious Jane” exist in the literature?', *Sex Roles*, 35: 461-88. 10.1007/BF01544132.
Webster, K., M. Recasens, V. Axelrod, and J. Baldridge. 2018. 'Mind the gap: A balanced corpus of gendered ambiguous pronouns', *Transactions of the Association for Computational Linguistics*, 6: 605-17.
Wehrmeyer, E. 2010. 'Animal characteristics in children’s literature: friends or scoundrels?', *Mousaion 28(3):85-100*.
Wikipedia contributors. 2020. 'Stereotypes of animals', Accessed 29.12.2020. https://en.wikipedia.org/w/index.php?title=Stereotypes_of_animals&oldid=994679581.
Wild, K., A. Kilgarriff, and D. Tugwell. 2012. 'The Oxford Children’s Corpus: Using a Children’s Corpus in Lexicography1', *International Journal of Lexicography*, 26: 190-218. 10.1093/ijl/ecs017.
Zhang, H., A. Sneyd, and M. Stevenson. 2020. 'Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs', *ArXiv*, abs/2010.02847.
Zito, J. 2018. 'Animal Protagonists in Children’s Literature'.

9 Appendix

9.1 Table with Target and Attribute Words

Table 8: Target and attribute word lists used for the WEAT.

ID	Attribute Words		Target Words	
	A	B	X	Y
Regular Vocabulary				
Gender bias				
G1	male, man, boy, brother, he, him, his, son, father, uncle, grandfather	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother	executive, management, professional, corporation, salary, office, business, career	home, parents, children, family, cousins, marriage, wedding, relatives
G2			math, algebra, geometry, calculus, equations, computation, numbers, addition	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
G3			science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
G4			precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
G5			power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser
Religious bias				
RL1	Jesus, Christian, gospel, church	Muhammad, Muslim, Koran, mosque	wonderful, best, superb, excellent	terrible, awful, worst, horrible
RL2	Jesus, Christian, gospel, church	Abraham, jew, torah, synagogue	wonderful, best, superb, excellent	terrible, awful, worst, horrible
RL3	Abraham, jew, torah, synagogue	Muhammad, Muslim, Koran, mosque	wonderful, best, superb, excellent	terrible, awful, worst, horrible
Age-related bias				
AG1	child, boy, girl, children, youth, lad	grandmother, granny, dowager, grandfather, grandpa, elder	excitement, delight, joyful, beautiful, cherish, fantastic, cheerful, magnificent	detest, poison, failure, pain, ugly, grief, nasty, despise
Animal-related stereotypes				
A1	aster, clover, hyacinth, marigold,	ant, caterpillar, flea, locust, spider, bedbug, centipede,	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond,	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault,

	poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia	fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil	gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation	disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison
A2	sheep, cattle, cows, mutton, woolly	coyote, wolf, werewolf	cute, guillable, sweet, innocent, love, peace, cheer, honest, lucky, happy, paradise, caress	cruel, evil, villainous, seductive, hungry, dangerous, lusting, disguise, sickness, accident, death, grief, poison
A3	bird, sparrow, crow, crane, chicken	fox	stupid, foolish, dull, naive, cowardly, frightened, scared, dumb, panicky, shy, worried	cruel, evil, villainous, seductive, hungry, dangerous, lusting, trick, wily, cunning, intelligent
A4	lion, sphinx, wildcat, puma, cougar, leo	mouse, mice, murine, hamster, rat, mole	strong, bright, pride, proud, brave, noble, royal, king, god, hero, ruler, hungry, majestic	tender, weak, quiet, silent, frail, mute, reticent, soundless, decrepit, vulnerable, infirm, sensible, nervous
A5	dog, puppy, bully, pup, doggy, hound, bitch, cur, mongrel	cat, kitten, bobcat, cheetah, cougar, jaguar, kitty, leopard, lion	faithful, nice, friend, smart, brave, happy, lovable, caring, peaceful, cute, polite, beautiful, cool, sly, charming, clever	selfish, wild, feral, mean, lazy, annoy, sad, careless, grumpy, boring, unfriendly, rude, evil, unlucky, gullible, depressed
Child-Specific Vocabulary				
Racial bias				
CR1	Meredith, Heather, Katie, Betsy	Latonya, Shavonn, Tashika, Ebony	good, nice, fun, happy	bad, mean, yucky, mad
Gender bias				
CG1	Ben, Peter, John, Tom	Alice, Jane Mary Wendy	addition, numbers, graph, math	read, books, story, letters
CG2	boy, he, him, his, man, father, brother, uncle, grandfather	girl, she, her, lady, women, mother, sister, aunt, grandmother	addition, numbers, graph, math	read, books, story, letters
Animal-related stereotypes				
CA1	daffodil, daisy, tulip, violet	bug, wasp, mosquito, roach	butterfly, heart, ice-cream, gift	wolf, witch, skull, injury

9.2 Sequence Diagram of the Bias Assessment Module

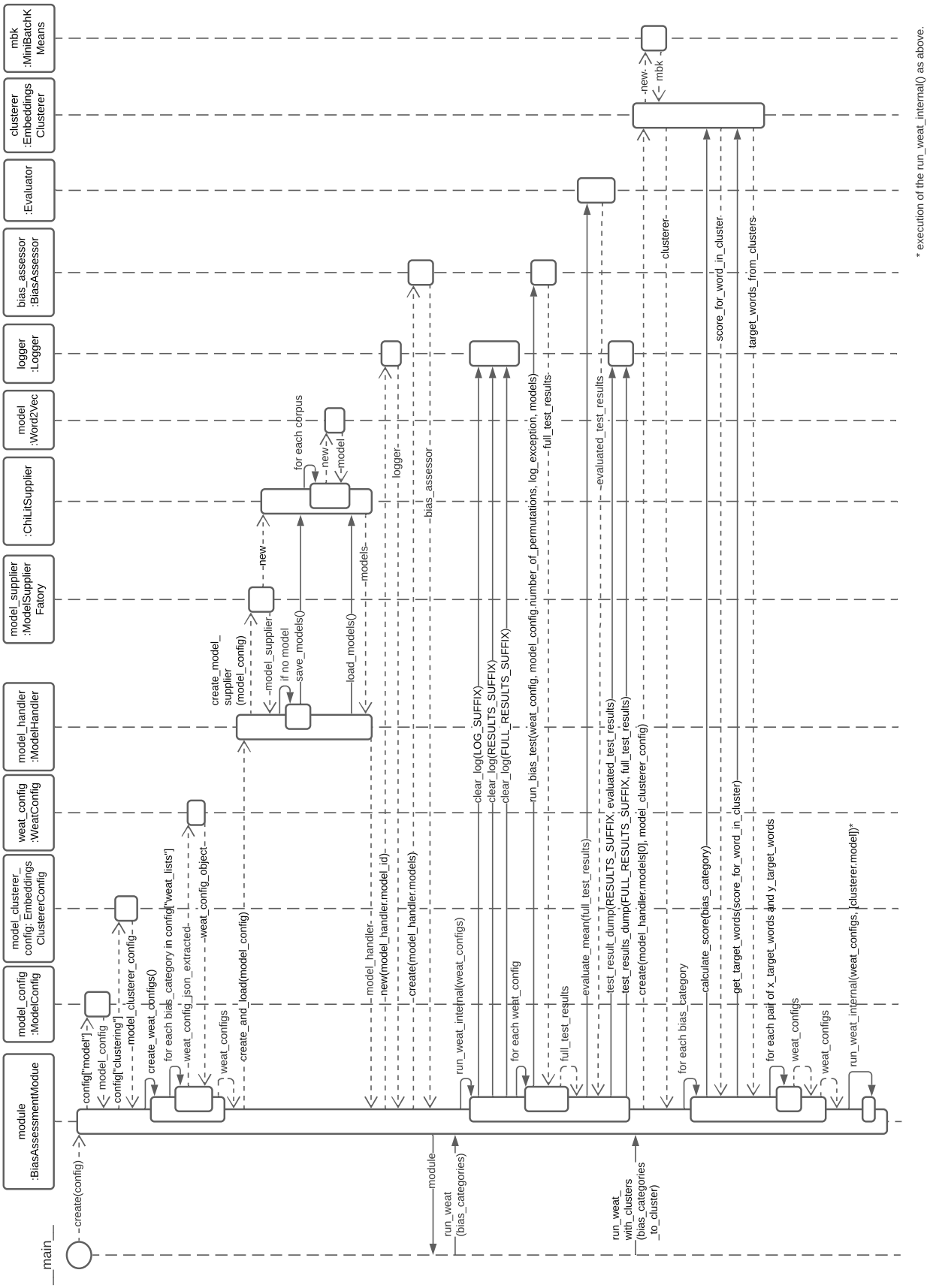

Figure 2: Simplified sequence diagram of the Bias Assessment Module for the example of supplying the ChiLit corpus.

9.3 Table with the WEAT results

Table 9: WEAT results, p-values in bold indicate statistically significant gender bias $p < 0.05$.

ID	ChiLit		ChiLit Small*		CLLIP		CLLIP Small*		CPB	
	p	d	p	d	p	d	p	d	p	d
context size window: 10 words										
G1	0.119	0.61	0.293	-0.18	0.325	0.23	0.349	-0.11	0.708	-1.73
G2	0.800	-0.52	-	-	0.172	0.61	0.363	0.06	0.500	-1.41
G3	0.369	-0.23	-	-	0.880	-0.71	-	-	-	-
G4	0.060	0.48	-	-	**0.007**	0.70	-	-	0.358	-0.08
G5	0.112	0.48	0.353	0.02	**0.044**	0.62	-	-	0.083	0.73
CG1	**0.000**	1.29	**0.000**	0.47	**0.000**	1.59	0.138	0.75	0.500	-1.41
CG2	0.416	-0.61	-	-	0.358	-0.26	-	-	**0.000**	-1.41
context size window: 30 words										
G1	**0.007**	1.16	0.290	-0.15	0.095	0.67	0.240	0.22	0.500	-1.15
G2	0.347	0.26	-	-	0.480	0.01	0.370	-0.44	0.500	-1.41
G3	0.326	-0.06	-	-	0.970	-1.04	0.355	-0.42		
G4	**0.033**	0.57	0.444	-0.13	**0.005**	0.73	0.292	0.11	0.300	-0.06
G5	0.149	0.41	0.353	0.04	0.070	0.54	0.260	0.39	0.316	-0.02
CG1	**0.000**	1.70	**0.050**	1.28	**0.058**	1.06	0.122	-0.04	**0.000**	1.41
CG2	0.250	0.68	-	-	0.526	-1.02	0.188	0.13	**0.000**	1.41

	ChiLit		ChiLit Small*		CLLIP		CLLIP Small*		CPB	
	p	d	p	d	p	d	p	d	p	d
context size window: 10 words										
RL1	**0.044**	1.02	-	-	**0.054**	0.93	-	-	-	-
RL2	0.207	0.33	-	-	0.315	-0.01	-	-	-	-
RL3	0.144	0.54	-	-	0.101	0.76	-	-	-	-
context size window: 30 words										
RL1	0.142	0.53	-	-	0.215	0.15	-	-	-	-
RL2	0.180	0.45	-	-	0.503	-0.66	-	-	-	-
RL3	0.388	-0.22	-	-	0.543	-0.86	-	-	-	-

	ChiLit		ChiLit Small*		CLLIP		CLLIP Small*		CPB	
	p	d	p	d	p	d	p	d	p	d
context size window: 10 words										
AG1	0.356	0.19	-	-	0.473	0.04	-	-	-	-
context size window: 30 words										
AG1	0.536	-0.05	-	-	0.300	0.28	0.160	-0.02	-	-

	ChiLit		ChiLit Small*		CLLIP		CLLIP Small*		CPB	
	p	d	p	d	p	d	p	d	p	d
context size window: 10 words										
A1	0.834	-0.29	0.292	0.17	0.422	0.06	-	-	0.596	-0.72
A2	0.378	0.14	-	-	0.606	-0.12	-	-	0.083	0.59
A3	0.643	-0.18	-	-	0.961	-0.79	-	-	0.208	1.00
A4	0.390	0.12	-	-	0.094	0.54	-	-	0.098	0.53
A5	0.725	-0.23	-	-	0.461	0.04	-	-	0.827	-0.60
CA1	0.330	-0.14	-	-	0.541	-0.94	-	-	**0.000**	1.41
context size window: 30 words										
A1	0.334	0.13	-	-	0.178	0.27	0.489	0.04	0.523	-1.08
A2	0.125	0.50	-	-	0.386	0.13	-	-	0.375	-0.81

A3	0.436	0.08	-	-	**0.049**	0.74	0.227	-0.41	0.166	0.95
A4	0.235	0.32	-	-	0.071	0.60	0.372	0.51	0.088	0.50
A5	0.745	-0.25	-	-	0.535	-0.03	0.339	0.06	0.684	-0.34
CA1	0.101	1.08	-	-	0.298	-0.21	-	-	**0.000**	-1.41

	ChiLit		ChiLit Small*		CLLIP		CLLIP Small*		CPB	
	p	d	p	d	p	d	p	d	p	d
context size window: 10 words										
CR1	**0.029**	1.16	**0.031**	1.12	0.513	-0.79	-	-	-	-
context size window: 30 words										
CR1	**0.058**	1.04	-	-	**0.036**	1.16				

* The results presented for ChiLit Small and CLLIP Small are averaged results of 10 randomly generated sub-corpora of ChiLit and CLLIP corpus respectively.

9.4 Table with the Out-Of-Vocabulary Words for each WEAT

Table 10: Number of out-of-vocabulary target and attribute words in the ChiLit, CLLIP and CPB embeddings.

Gender bias

			G1		G2		G3		G4		G5				CG1				CG2	
	A	B	X	Y	X	Y	X	Y	X	Y	X	Y	A	B	X	Y	A	B	X	Y
Total	11	11	8	8	8	8	8	8	25	25	15	15	4	4	4	4	9	9	4	4
ChiLit	0	0	0	0	2	2	4	2	4	2	1	2	0	0	2	0	0	0	2	0
CLLIP	0	0	0	0	3	1	2	1	2	1	0	0	0	0	1	0	0	0	1	0
CPB	1	2	6	2	7	7			22	17	12	10	1	2	3	0	0	0	3	0

Religious bias

			RL1				RL2				RL3		
	A	B	X	Y	A	B	X	Y	A	B	X	Y
Total	4	4	4	4	4	4	4	4	4	4	4	4
ChiLit	2	3	0	0	2	3	0	0	3	3	0	0
CLLIP	2	3	0	0	2	1	0	0	1	3	0	0
CPB												

Age-related bias

			AG1	
	A	B	X	Y
Total	6	6	8	8
ChiLit	0	0	0	0
CLLIP	0	1	0	0
CPB				

Animal-related stereotypes

			A1				A2				A3				A4			
	A	B	X	Y	A	B	X	Y	A	B	X	Y	A	B	X	Y
Total	25	25	25	25	5	3	12	13	5	1	11	11	6	6	13	13
ChiLit	7	9	1	2	0	1	1	1	0	0	2	1	2	2	0	2
CLLIP	11	6	2	1	0	0	1	1	0	0	0	1	3	1	0	1
CPB	21	18	13	21	2	2	6	11	2	0	6	9	5	5	7	9

			A5				CA1	
	A	B	X	Y	A	B	X	Y
Total	9	9	16	16	4	4	4	4
ChiLit	2	3	0	2	0	1	1	0
CLLIP	2	4	0	0	0	0	1	0
CPB	7	6	6	11	3	3	1	3

Racial Bias

			CR1	
	A	B	X	Y
Total	4	4	4	4
ChiLit	0	3	0	1
CLLIP	0	3	0	0
CPB				

9.5 Changelog

commit 6df62ab
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Jan 25 11:10:25 2021 +0100

- documentation

commit 511ace0
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Jan 25 11:06:25 2021 +0100

- documentation

commit 1b47e4e
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Jan 24 23:49:30 2021 +0100

- documentation

commit c6ae803
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Jan 24 11:05:16 2021 +0100

- added books for teenagers to the CLLIP Corpus
- added documentation for ChiLitSupplier and CLLIPSupplier

commit f2f1b1d
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sat Jan 23 23:57:00 2021 +0100

- created documentation of classes and their main functions
- changed the corpus size constant for the creation of the sub-corpora
- minor refactorings

commit 382f0d6
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Jan 11 00:32:28 2021 +0100

- function _run_weat renamed to _run_weat_internal
- replaced test_results_dump for clearing test results by clear_log function in the _run_weat_internal
- bug fix for performing clustering: words that are not in the vocabulary are now filtered out

- added new argument 'mode' to start the Bias Assessment Module and implemented the functionality in the main method
- added the mode parameter to the config.json function

commit d93227c
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Fri Jan 8 22:46:03 2021 +0100

- moved all supplier classes to the folder "supplier"
- fixed bugs after the major refactoring of the application
  - fixed saving models for GoogleNews and GAP corpus in the cache folder
  - replaced all occurrences of model_config calls in json format by properties of the ModelConfig
  - config_to_id function has been made compatible with the ModelConfig object
  - fixed extracting weat config properties from the json file in the create_weat_configs function in the BiasAssessmentModule class
  - added Exception handling in the BiasAssessmentModule and BiasAssessor for PreconditionErrors
  - replaced all occurrences of cluster_config calls in json format by properties of the ClusterConfig
  - added a precondition in evaluate_mean function in the Evaluator class, as the category test results are empty in some cases.
  - added name of the file where the model is stored as a parameter to the _config.jsons of the GAP and GoogleNews corpus
- defined and integrated command line arguments to start the application
- added an additional parameter to config.json which is used to determine whether clustering should be executed or not

commit fc5382e
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Fri Jan 8 01:17:33 2021 +0100

- restructured config.json - removed unnecessary depth level in weat_lists
- added and implemented three classes EmbeddingsClustererConfig, ModelConfig, WeatConfig to get structured data from config.json
- refactored the WeatTester class:
  - moved run_weat_test function to the BiasAssessmentModule class
  - moved and refactored start_clustering function to the EmbeddingsClusterer class
  - deleted the WeatTester class
- added a list with the bias categories that can be used for clustering to the main method
- added the option to run the WEAT with clusters from the main method
- refactored BiasAssessmentModule:
  - added new members
  - implemented method to create weat config objects from the config.json
  - major refactoring of the run_weat_test function. Now it is possible to use internal _run_weat function for both normal WEATs and WEATs with clusters.
  - moved logging and creating dumps of the results to the new Logger class

In the BiasAssessor class:
- removed config from the BiasAssesor class and replaced it with using the properties of the WeatConfig class
- removed bias_test_for_clusters, a method to run the WEAT with clusters, as it now can be run by the BiasAssementModule class
- refactored run_bias_test function > added possibility to pass the model and the number of permutations

commit d5695fd
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Jan 7 15:35:09 2021 +0100

- review and refactoring of variables representing target and attribute words in BiasAssessor and in EmbeddingsClusterer
- added functionality to recursively read files in a corpus
- extracted functionality to start clustering in a function

commit a25e445
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Wed Jan 6 01:50:19 2021 +0100

- started to add user input parameters in WeatTester
- evaluation of user input in the BiasAssessmentModule class
- moved initialisation of the EmbeddingsClusterer from the BiasAssessmentModule to the WeatTester class
- EmbeddingsClusterer class tested with new model structure and fixed minor bugs
- bug fix for the get_files function, as it was loading hidden files beginning with "."

commit 32c901a
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Jan 4 22:52:51 2021 +0100

- moved the loading and saving of the model from the class ChiLitSupplier to the abstract class ModelSupplier
- implemented algorithms for loading and saving the model in ModelAndCorpusSupplier
- added directory creation for model storage
- implemented model loading and saving for the GoogleNewsSupplier, GAPSupplier, CPBCSupplier and CLLICSupplier classes

commit a8f96e8
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Jan 4 02:09:04 2021 +0100

- separated the process of generating multiple sub-corpora from a corpus and loading corpus data

- moved algorithm for generating multiple sub-corpora from the ChiLitSupplier class to the abstract parent class CorpusSupplier, as the algorithm is the same for other concrete CorpusSuppliers
- moved the constant CORPUS_SIZE from ChiLitSupplier to CorpusSupplier
- added two abstract methods to the class CorpusSupplier for loading a single corpus or multiple corpora which are implemented in respective concrete CorpusSuppliers

commit 3928cac
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Jan 3 23:59:32 2021 +0100

- refactored BiasAssessmentModule.py:
  - removed test_results property and moved deleting old entries in the results file to the WeatTester class
  - refactored test_result_dump function > moved functionalities to test_results_dump function; added internal multiple model evaluation to the test test_results_dump function
- renamed variables in start_bias_test in BiasAsessor
- refactored and improved _load_data function in ChiLitSupplier.py for multiple corpora support
- refactored and improved calculation of mean data in the class Evaluator
- implemented support of GAP Corpus model
- added setters to the TestResult class
- multiple file support for test results in WeatTester.py

commit e12af4f
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Jan 3 00:16:32 2021 +0100

work-in-progress 2
- implemented functionality to randomly generate the number of sub-corpora specified in config.json for ChiLit Corpus
- added functions in the Evaluator.py class to calculate mean p-value, mean cohen's d, mean number of permutations and mean total time

commit 9bdf033
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sat Jan 2 01:15:38 2021 +0100

Work-in-progress
- Extensive refactoring and remodeling in order to add the functionality for generating sub-corpora of large corpora:
  - start_bias_test() in BiasAssessor extended with functionality to run WEAT for multiple models
  - extended CorpusSupplier with model functionality and renamed to ModelAndCorpusSupplier
  - created the ModelSupplier abstact class for the persistance of the models

- each corpus implements functionality of the new base type ModelAndCorpusSupplier, which in turn inherits from the abstract classes ModelSupplier and CorpusSupplier
- added GoogleNewsSupplier and GAPSuplier classes that will load the corresponding corpora
- load() function from the class ModelHander restructured > functionalities outsourced to the ModelSupplier and CorpusSupplier classes
- renamed corpus_type to model_type in the config.json-file for each corpus
- changed mapping names for model_type(s) in ModelAndCorpusSupplier
- added Evaluator class responsible for evaluating WEAT results from multiple models and averaging the p-value and cohen's d

commit b23963c
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Dec 31 20:37:09 2020 +0100

- added removal of the book author and title for CorpusSupplierChiLit

commit 1659d4a
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Dec 31 15:35:05 2020 +0100

- added Children's Picture Books Corpus with corresponding _config.json
- created a CorpusSupplierCPBC class:
- implemented function load_data(), which removes book titles and authors from the original text file and preprocesses the cleaned data with gensim
- added CorpusSupplierCPBC the CorpusSupplierFactory

commit 1bcb189
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Dec 3 23:42:07 2020 +0100

- implemented functionality to balance the sets of attribute and target words if the words from the sets are missing in the vocabulary of the corpus.

commit 0703019
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Wed Dec 2 22:54:54 2020 +0100

- added new categories wolf_sheep, fox_bird, lion_mouse with corresponding target and attribute words to the weat_list in config.json
- added new categories to the bias_categories list in the WeatTester class

commit 6a6b3d1
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Tue Dec 1 23:49:00 2020 +0100

- added categories gender.b1, gender.b2, gender.b3, gender.b4, gender.b5 after Chaloner, K., & Maldonado, A. (2019). Measuring Gender Bias in Word Embeddings with corresponding target and attribute words to the weat_list in config.json
- added category flower_vs_insects after Caliskan, A., Bryson, J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases.
- moved first drafts for test executions and error handling from the run() method in the BiasAssessmentModule class to the WeatTester class
- implemented a WeatTester class which:
  - starts the weat tests for each bias category
  - generalizes error handling for all tests if one of the arrays with attributes or target words is empty
- implemented a BiasAssessorException class to handle errors that occur during the execution of the weat tests
- added fasttext and KeyedVectors libraries for testing GAP and Google News corpora in the ModelHandler class
- implemented error handling in the BiasAssessor class + added function to print the progress of running the permutation tests
- moved the creation and loading of the model and the creation of the BiasAssessor object to the constructor of the BiasAssessmentModule from the run() method

commit d0770c6
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Nov 30 23:57:53 2020 +0100

- added categories race, gender_math, religion_with_names, religion_christianity_islam, religion_christianity_judaism, religion_judaism_islam with corresponding target and attribute words to the weat_list in config.json
- added a help function one_is_empty in Utils which checks if one of the arrays with attributes or target words is empty
- first draft for exception handling if one of the arrays with attributes or target words is empty
- added preconditions for the bias_test function in the BiasAssessor class
- first draft of tests for each bias category

commit 5038f74
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Nov 29 23:38:04 2020 +0100

- added category "animals" and "cat_dog" with corresponding target and attribute words to the weat_list in config.json
- extracted configuration to start the bias test from the BiasAssessmentModule class in a start_bias_test() function in the BiasAssessor class
- minor bug fixes

commit 342e8c2
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Nov 23 23:25:31 2020 +0100

- added the first draft description of the Bias Assessment Module in README.md
- refactored the project
  - moved all relevant classes from the root directory to bias_assessment_module
  - deleted or archived files that are no longer needed
- changed output formatting of test results using named arguments
- added two new parameters to the config file of ChiLit_Corpus_Small - corpus_length and single_text_length and used them in the CorpusSupplierChiLitSmall class
- added an additional parameter "config" to the constructor of the CorpusSupplier class and implemented it in all child classes

commit 07e066d
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Nov 22 23:51:43 2020 +0100

- implemented a new class CorpusSupplierChiLitSmall to create and process a subset of ChiLit corpus
- added a _config.json in data/ChiLit_Corpus_Small, which maps the subset of the ChiLit corpus to the class CorpusSupplierChiLitSmall
- refactored functions and parameters in BiasAssessmentModule and BiasAssessor in order to make them more general
- simplified test_result_dump() function in the BiasAssessmentModule: the dump format is now clearer
- amended and refactored functions for calculating the score of a word and getting target words in the EmbeddingsClusterer class
- changed the data structures for transferring scores for words in clusters and for storing the target words
- bugfix for creating fixed size clusters (added an additional parameter to config.json and used it in the EmbeddingsClusterer class)
- added README.md

commit f68b1d8
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Nov 22 01:23:27 2020 +0100

- added configuration parameters of the k-means++ algorithm to config.json
- further adapted the clustering approach to identify new target words for gender bias:
  - based on the object-oriented approach, implemented the creation and processing of the EmbeddingsClusterer object in the BiasAssessmentModule class
  - adapted the functions provided by Chaloner & Maldonado(2019) for clustering and processing the clustering results: the functions calculate_score() and the function for getting m and f target words have been simplified
  - implemented a new function in BiasAssessor that calls bias_test() with the parameters found as a result of clustering
  - adapted the function bias_test() so that it can now be used with predefined target words as well as with target words found with the k-means++ algorithm

- modified the test_result_dump() function to append new results to a file instead of overwriting the file each call

commit c1da7d1
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Fri Nov 20 23:52:02 2020 +0100

- implemented a prettify_test_result method to print the selected results to standard output
- created new class Utils and moved the filter_words function from Bias Assessor to the class Utils
- started to adapt the clustering approach to detect new target words for gender bias

commit 5a17bbd
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Nov 19 23:58:45 2020 +0100

- implemented functionality to save the results of the weat-tests in a file:
- implemented the class TestResult which represents the result of a test, and used it in the function bias_test()
- implemented function test_result_dump() for saving and formatting several weat-test results
- extended the function filter_words() in the BiasAssessor class to get the filtered_words() in addition to the the final_words
- added the time metric to the test procedure

commit 8db244f
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Wed Nov 18 23:36:40 2020 +0100

- added a config.json file containing all information needed to train and cache the word2vec model and the weat_lists with target and attribute words for different bias categories
- moved attribute and target words from txt files to the config.json file
- extracted variable data from methods of the ModelHandler class into the config.json
- implemented run() in BiasAssessmentModule which loads the config.json, creates a ModelHandler and BiasAssessor objects, and executes bias_tests
- create() method implemented in BiasAssessor class
- some methods in the BiasAssessor class are static now
- adjusted caching mechanism in the ModelHandler class:
- implemented _config_to_id() method that dynamically generates an id for a corpus based on the model configuration from config.json
- If there is no model with the generated id in data/cache, the new model is trained and stored under the generated id. If a model already exists, it is loaded from data/cache.

commit f068b35
Author: vlebedynska <v.lebedynska@gmail.com>

Date: Tue Nov 17 23:55:51 2020 +0100

    - added lists of target and attribute words needed to measure gender bias using the WEAT hypothesis test
    - started implementing BiasAssessmentModule class as a main module class of the Bias Assessment Module
    - moved functionalities for the WEAT hypothesis test from preprocessing_module.py to the BiasAssessor class
    - made the CorpusSupplier class abstract using ABC lib and implemented a constructor there which is valid for subclasses
    - extracted the function for loading corpus files from CorpusSupplierChiLit and CorpusSupplierCLLIC into CorpusSupplier base class by implementing the function get_files(). It's a generator function which iterates over file paths. Data from files is loaded in CorpusSupplierChiLit and CorpusSupplierCLLIC while iterating through the get_files() generator.
    - implemented a mechanism in the CorpusSupplierCLLIC class which parses CLLIC corpus data (xml files) in plain text.
    - implemented a new class CorpusSupplierFactory based on the factory pattern and moved the functionality for providing the corpus from the ModelHandler class there
    - implemented in ModelHandler a simple database for saving and loading w2v models

commit 2fa272b
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Nov 16 23:32:15 2020 +0100

    - created _config.json files for each corpus. Each config file contains information about corpus type, which is necessary in order to set the correct type of CorpusSupplier in ModelHandler
    - in the ModelHandler class, implemented a function create_corpus_supplier() which dynamically creates a corpus supplier depending on the information contained in _config.json
    - changed CorpusSupplier to an abstract class
    - let CorpusSupplierChiLit and CorpusSupplierCLLIC inherit from CorpusSupplier
    - in CorpusSupplierChiLit implemented a constructor and a function to load document by document from the corpus folder
    - extracted functionalities to load and save the model from preprocessing_module.py to the ModelHandler class
    - added ModelHandler initialisation in preprocessing_module.py
    - new class BiasAssessmentModule added

commit 88bbbfe
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Nov 15 23:30:31 2020 +0100

    - added data to the project: ChiLit with children's literature from the 19th century and CLLIP corpus with fiction books written for children in the 20th century
    - created the basic class structure for the Bias Assessment Module

- added justTheWords.xsl, display.xsl and oneWordPerLine.xsl stylesheets provided by BNC
- modified justTheWords.xsl to display one sentence per line

commit 561b0f9
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Thu Nov 12 23:57:20 2020 +0100

- added test files from the CHILDES corpus and the British National Corpus
- added data for testing unsupervised detection of new bias categories
- added bert.py for testing the implementation of the BERT model
- implemented helper functions in helper.py to extract references to texts written for a child audience from the British National Corpus (rebuilding of the CLLIP Corpus)
- added auto-generated file books_extracted.txt as the result of the extraction processes to git
- implemented the k-means++ algorithm in order to extract new bias categories from word embeddings trained on children's books and added functionality to load, build and process clusters in preprocessing_module.py after Kaytlin Chaloner and Alfredo Maldonado (2019). Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories (https://github.com/alfredomg/GeBNLP2019)
- started to implement a normalizer for crawling BNC corpus and GLARE 19th Century Children’s Literature Corpus and for preprocessing data extracted from the corpora

commit ec89f37
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Mon Oct 26 22:58:46 2020 +0100

- extracted functionalities for loading data, training and saving the model in load and save functions
- implemented functionalities for gender bias testing provided by Kaytlin Chaloner and Alfredo Maldonado (2019). Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing. Florence.
- added logging for better interpretation of test results

commit 53a1ed5
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sun Oct 25 23:26:43 2020 +0100

- test sources added (children's books training corpus from The Children's Book Test)
- implemented test functions for:
- reading the text file and splitting the text by _BOOK_TITLE_ in separate documents
- simple pre-processing of the document
- training the Word2Vec model
- outputting the cosine similarity and most similar words
- added lists of exemplary target words and attribute words as preparation for calculating the difference between target words in relation to attribute words

commit 278d9ea
Author: vlebedynska <v.lebedynska@gmail.com>
Date: Sat Oct 17 21:27:25 2020 +0200

    Initial commit

10 Statement of Independent Work

Hiermit versichere ich an Eides Statt, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken und Quellen, einschließlich der Quellen aus dem Internet, entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht.

Dasselbe gilt sinngemäß für Tabellen, Karten und Abbildungen. Diese Arbeit habe ich in gleicher oder ähnlicher Form oder auszugsweise nicht im Rahmen einer anderen Prüfung eingereicht.

Köln, 25.01.2021 Unterschrift Viktorya Olari