EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D6.11: Final report on gender bias in content creation (T6.4)

Executive summary

This deliverable presents the work conducted within EMBEDDIA on gender bias in content creation systems, and particularly on gender bias in word embeddings, natural language generation systems, and automated journalism. The main findings and contributions are the following. First, our study of various Slovene and Croatian word embeddings that focuses on vocabulary related to professional occupations finds that there is very little bias when searching for the masculine occupations, given the feminine equivalent – however, in the opposite direction, the bias is very prominent. Second, we present a new method for de-biasing word embeddings via corpus transformation, and find that it removes direct gender bias. Third, we present a study on gender biases in the GPT-2 language model, and find that, when conditioned on gender-neutral words that describe occupations, the model is strongly biased towards generating male gendered keywords. The report also includes a discussion on using natural language generation tools in journalism, and concludes with suggestions for future work.

Partner in charge: UH

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D6.11**
Deliverable full title:	**Final report on gender bias in content creation**
Deliverable short title:	**Final gender bias**
Document identifier:	**EMBEDDIA-D611-FinalGenderBias-T64-submitted**
Lead partner short name:	**UH**
Report version:	**submitted**
Report submission date:	**31/12/2021**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Michael Mathioudakis (UH), Leo Leppänen (UH), Senja Pollak (JSI), Matthew Purver (QMUL)**
Co-author(s):	**Matej Ulčar (UL)**
Status:	**__ draft, __ final, _x_ submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
12/11/2021	v00	Michael Mathioudakis (UH)	Outline.
24/11/2021	v01	Leo Leppänen (UH)	Automated journalism.
24/11/2021	v02	Matej Ulčar(UL), Senja Pollak (JSI)	Biases in word embeddings.
29/11/2021	v03	Michael Mathioudakis (UH)	Biases in NLG models.
10/12/2021	v04	Matthew Purver (QMUL)	Bias mitigation.
12/12/2021	v05	Matej Ulčar(UL), Senja Pollak (JSI)	Biases in word embeddings.
17/12/2021	v06	Michael Mathioudakis (UH)	Draft for review.
20/12/2021	v07	Marko Robnik-Šikonja (UL)	Internal review.
20/12/2021	v08	Antoine Doucet (ULR)	Internal review.
21/12/2021	v08	Michael Mathioudakis (UH)	Addresses some of the comments of the internal review.
22/12/2021	v09	Nada Lavrač (JSI)	Quality control.
24/12/2021	final	Michael Mathioudakis (UH)	Report finalized.
27/12/2021	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

CSLS	Cross-Domain Similarity Local Scaling
EC	European Commission
EU	European Union
DoA	Description of Action

1 Introduction

In this deliverable, D6.11 of task T6.4, we describe the work related to gender bias in language that was conducted by the members of the EMBEDDIA consortium. Recall that the goal of Task 6.4 is to "propose means to avoid gender and other biases in news media contents creation". While in D6.1, the first deliverable of the task, we already provided recommendations to avoid such biases, in this deliverable we describe our research within EMBEDDIA that largely deals with technical issues related to the gender bias in content creation. The work described below is partly related to work packages WP1–2, as some of the work is related to word embeddings (Section 2), as well as to work packages WP3–5, because some of the work is related to content generation (Sections 3–4).

The structure of the report is as follows. The first part is devoted to gender bias in word embeddings (Section 2). Specifically, we present works that study gender bias in Slovene and Croatian word embeddings, particularly in relation to what professions are typically associated with each gender (Sections 2.1–2.2), as well as a method to mitigate the gender bias in embeddings (Section 2.3). In the second part, we present works that study gender bias in the context of Natural Language Generation (NLG) tools (Sections 3–4): Section 3 presents a study of gender biases in the `GPT-2` language model, and Section 4 discusses the promise and challenges of using NLG tools in automated journalism. Finally, Section 5 concludes with ideas for future research directions.

2 Gender Bias in Word Embeddings

Deep neural networks are the predominant learning method for many text analytic tasks. On their input they expect that words are encoded with numerical vectors, called word embeddings. A common procedure to build word embeddings is to train a neural network on one or more semantic text classification tasks and then take the weights of the trained neural network as a representation for each text unit (word, n-gram, sentence, or document). The labels required for training such a classifier come from huge corpora of available texts. Typically, they reflect word co-occurrence, like predicting the next or previous word in a sequence or filling in missing words but may be extended with other related tasks, such as sentence entailment. The positive instances used for training are obtained from texts in the used corpora, while the negative instances are mainly obtained with negative sampling (sampling from instances that are highly unlikely related).

The relations between words are expressed in the geometry of the embedded vector space: semantically related embeddings lie close in the vector space and are arranged in similar directions. This enables the study of relations beyond superficial similarities between words, e.g., through analogies (Mikolov et al., 2013). Biases in word embeddings manifest through semantic associations and consequent proximities in the vector space (Mikolov et al., 2013), and therefore can reflect biases present in human language, including the gender bias (Bolukbasi et al., 2016). We can thus measure gender bias using the word analogy task.

We analyse gender bias in language models, where we experiment with Slovene and Croatian language models thought the prism of occupational analogies. In Section 2.1, we present the experiments on quantitative evaluation of gender bias in several Slovene and Croatian language models, while in Section 2.2 we present how this type of experiments can also serve as the basis of interdisciplinary work including more qualitative interpretation. In Section 2.3, we provide a method to de-bias word embeddings via corpus transformation.

2.1 Gender bias evaluation in Slovene and Croatian word embeddings

The work presented in Section 2.1 is described in detail in the papers by Ulčar et al. (2021) and Supej et al. (2020), attached as Appendix A and Appendix B, respectively.

We analyzed several Slovene and Croatian word embeddings, evaluating the analogy relations between the equivalent masculine and feminine nouns for occupations.

2.1.1 Datasets

We compiled two lists of occupations, one for Slovene and one for Croatian. The Slovene list is based on the Standard Classification of Occupations (Vlada RS, 1997), based on the International Standard Classification of Occupations. We limited our evaluation to single-word occupations, as multi-word expressions are less suitable for this task due to their specificity and length (e.g., metallurgical crane operator). We removed the occupations which appear less than 500 times in the corpus of written standard Slovene Gigafida 2.0 (Krek et al., 2016). We manually added synonyms of occupations if the synonym is more established than the standard form. If the standard classification was missing either a male or female variant of an occupation, we manually added it, provided it appears in the Gigafida corpus. The final list contains 234 occupation pairs and was made publicly available.[1]

The Croatian occupation list is based on the word analogy dataset by Svoboda & Beliga (2018) and on the ESCO[2] (European Skills, Competences, Qualification and Occupations) list. We combined the two lists and removed all the multi-word occupations. The combined filtered list contains 375 occupation pairs.

In the bias analysis we tested several Slovene and Croatian word embeddings. For Slovene, we used 256-dimensional word2vec embeddings, trained for the needs of the Kontekst.io portal [3] (Plahuta, 2020), 1024-dimensional ELMo embeddings (Ulčar & Robnik-Šikonja, 2020) for each ELMo layer (centroid vectors calculated on Slovene Wikipedia for 200,000 most common words) and several fastText embeddings: 300-dimensional embeddings from the fastText.cc portal, 100-dimensional CLARIN.SI-embed.sl embeddings (Ljubešić & Erjavec, 2018), 100-dimensional word and lemma embeddings from the Sketch Engine, and 100- and 300-dimensional embeddings we trained on Gigafida 2.0 corpus.

For Croatian, we used the following fastText embeddings: 300-dimensional embeddings from the fastText.cc portal, 100-dimensional CLARIN.SI-embed.hr embeddings (Ljubešić, 2018), 100- and 300-dimensional embeddings, trained in the EMBEDDIA project.

2.1.2 Methodology

We measured the gender bias in word embeddings by evaluating the analogy relation "man is to male occupation what woman is to female occupation". Specifically, for every masculine occupation noun O_m, we calculated the vector:

$$v(d) = v(O_m) - v(m) + v(f),$$

where $v(m)$ is the male vector, and $v(f)$ is the female vector. If there were no gender biases in the embeddings model, $v(d)$ would be very similar to $v(O_f)$, that is, the feminine noun equivalent of O_m. For every vector $v(d)$, we found the N closest word vectors, measured with cosine similarity metric. If any $v(O_f^{lemma})$ is among these N word vectors, we counted the analogy as correctly solved. Here, O_f^{lemma} are all the words that have the same lemma as O_f. The average accuracy of this procedure is commonly

[1] http://hdl.handle.net/11356/1347

[2] https://ec.europa.eu/esco/portal

[3] https://kontekst.io/ is an associative dictionary of Slovene, Croatian, and Serbian, based on word embeddings. Slovene word embeddings were trained using word2vec on around 15 Gb of text (academic, news, books etc.).

called Precision@N or P@N. We repeated the same process by swapping the genders, such that for every feminine occupation noun O_f, we calculated the vector:

$$v(d) = v(O_f) - v(f) + v(m).$$

Again, we found the closest N word vectors for each $v(d)$, and counted the analogy as correctly solved if any $v(O_m^{lemma})$ was among them, where O_m^{lemma} are all the words that have the same lemma as O_m.

For calculating male and female vectors, $v(m)$ and $v(f)$, respectively, we used two approaches. In the first approach $v(m)$ is equivalent to the embedding of the word 'man' and $v(f)$ is equivalent to the embedding of the word 'woman'. In the second approach, the difference between $v(m)$ and $v(f)$ is calculated as the average difference between vectors of word pairs, which have a natural male and female counterparts, for example 'man' and 'woman', 'boy' and 'girl', 'brother' and 'sister', 'father' and 'mother', 'he' and 'she', 'son' and 'daughter'.

2.1.3 Results

We present the results for all the embeddings presented in Section 2.1.1, using the second approach for male and female vectors described in Section 2.1.2, i.e. the average of several inherently male and female words. The results for Slovene embeddings are shown in Table 1 and the results for Croatian embeddings in Table 2. We use the P@N measure, where N equals 1, 5, or 10. Some of the occupations from our list are not covered by all word embeddings, i.e. there is no word vector for them. Any example where the searched-for word is not among the top N closest words is counted as incorrect, even if the searched-for word does not appear in the embeddings. In cases where the embeddings do not cover the input occupation, and we cannot calculate the vector v(d), we dismiss such examples so that they do not affect the final result.

		f input			*m* input		
Slovene embeddings	dimensions	P@1	P@5	P@10	P@1	P@5	P@10
ELMo Embeddia	1024D l0	0.907	0.933	0.947	0.370	0.398	0.403
	1024D l1	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l2	0.880	0.933	0.933	0.376	0.398	0.398
fastText.cc	300D	0.613	0.884	0.948	0.655	0.755	0.764
fastText Embeddia	100D	0.906	0.971	0.976	0.677	0.720	0.724
	300D	**0.947**	**0.976**	**0.982**	0.685	0.720	0.724
fastText CLARIN.SI-embed.sl	100D	0.839	0.940	0.950	**0.761**	**0.880**	**0.902**
fastText Sketch Engine (word)	100D	0.930	0.962	0.973	0.725	0.781	0.785
fastText Sketch Engine (lemma)	100D	0.673	0.931	0.960	0.598	0.786	0.821
word2vec Kontekst.io	256D	0.679	0.853	0.872	0.407	0.550	0.593

Table 1: Results for all Slovenian embeddings for each approach, where we have a feminine word for occupation on the input (f input), and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (m input), and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

The results show a great variance between different embeddings. In both Slovene and Croatian, our fastText 300-dimensional embeddings show the least bias when feminine word for occupation is presented on the input and we search for the equivalent masculine term. When masculine occupation is presented on the input and we search for the equivalent feminine term, the fastText embeddings from CLARIN.SI portal show the least bias.

The best performing embeddings show very little bias when searching for the masculine occupations, given the feminine equivalent. In the opposite direction, the results are much worse. This can be explained largely by the fact that much fewer feminine occupation nouns are covered by the embedding

Croatian embeddings	dimensions	*f* input P@1	*f* input P@5	*f* input P@10	*m* input P@1	*m* input P@5	*m* input P@10
fastText.cc	300D	0.731	0.939	0.954	0.546	0.637	0.644
fastText Embeddia	100D	0.905	0.941	0.968	0.625	0.666	0.672
	300D	**0.923**	**0.982**	**0.986**	0.631	0.675	0.678
fastText CLARIN.SI-embed.hr (word)	100D	0.907	0.930	0.944	**0.673**	**0.746**	**0.754**
fastText CLARIN.SI-embed.hr (lemma)	100D	0.244	0.678	0.826	0.266	0.521	0.588

Table 2: Results for all Croatian embeddings for each approach, where we have a feminine word for occupation on the input (f input), and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (m input), and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

models, as shown in Table 3. Except for ELMo embeddings, which were limited to 200,000 most common words, more than 97% of all Slovene masculine occupation nouns are covered by each of the embeddings. However, only one model covers more than 90% of Slovene feminine occupation nouns, most covering less than 75%. For Croatian occupations, the numbers are lower for both genders, but the difference remains similar.

Slovene embeddings	m	f	Croatian embeddings	m	f
ELMo	0.774	0.321			
fastText cc	0.979	0.739	fastText cc	0.848	0.527
fastText Embeddia	0.991	0.726	fastText Embeddia	0.856	0.594
fastText CLARIN.SI-embedd.sl	1.000	0.932	fastText CLARIN.SI-embedd.hr (word)	0.914	0.722
fastText Sketch Engine (lemma)	1.000	0.863	fastText CLARIN.si-embedd.hr (lemma)	0.955	0.722
fastText Sketch Engine (word)	0.996	0.791			
word2vec Kontekst.io	0.987	0.667			

Table 3: Coverage of male (m) and female (f) occupations from the list in different embeddings as a ratio between covered occupations and all occupations.

Masculine occupations that do not appear in the embeddings are typically occupations associated with women (e.g., male variants of *seamstress* and *cosmetician*, in Slovene *šiviljec* and *kozmetik*, respectively). Likewise, feminine occupations not present in the embeddings are traditionally male occupations (e.g., embedding models do not contain female variants of occupations like *auto mechanic* and *carpenter* (in Slovene *avtomehaničarka* and *tesarka*, respectively), or occupations that have been culturally taken up exclusively by men, e.g., *nadškof* (en. *archbishop*). Poor representation of female occupations can also be attributed to other factors. Zhao, Wang, et al. (2018) report that the mentions referring to men are more likely to contain a job title compared to female mentions.

We observed that certain words (especially female occupations) frequently appear among the results despite being semantically unrelated to the input occupation. Several analogy results (especially in the case of a typical male occupation on the input) are unrelated to the input occupation (e.g., *bolničarka* [en. $nurse_F$] is the first result of the analogy *moški:rudar :: ženska:x* [en. *man:miner :: woman:x*] and *šivilja* [en. *seamstress*] is the first result of the analogy *moški:avtomehanik :: ženska:x* [en. *man:auto mechanic :: woman:x*] in the Slovene model named *fastText Embeddia 100D*). One explanation is that certain word vectors are more *central* than the others and, therefore, the closest neighbours of many other words. To check if this explanation is true, instead of the cosine similarity measure, we used the CSLS measure (Conneau et al., 2018) that considers the shared distances of N closest neighbours. We have found that the distribution of the most common words in the results is more uniform when using the CSLS measure. However, the overall precision, reported in Table 1 and Table 2 is worse when using the CSLS measure compared to the cosine similarity measure.

2.2 Gender bias and society

The work presented in Section 2.2 is described in full in papers by Supej et al. (2019), attached here as Appendix C.

In our socio-linguistic study Supej et al. (2019), we focused on a single embeddings model (Slovene word2vec), but put more attention to the qualitative interpretation of results.

The occupational pairs were similar to the ones in Section 2.1 with the exception that we selected only two groups of occupations where men and women had the highest quantitative hourly wage difference: (1) Legislators, senior officials, managers and (2) Experts, but also included occupations from the group with the smallest difference, i.e. Officials (Eurostat & SURS, 2018).

The quantitative results were very similar to those in the previous study (ranging from 71% P@1 for female seed words to 98% P@10 for male seed words). The complete detailed results are provided in the article in Appendix C.

From the qualitative interpretation, the analogy *secretaryF : bossM* clearly stands out as an example, where the gender analogy expresses a hierarchical relation, and therefore reflects societal inequalities. Another interesting examples is that the candidates for female analogue to 'dancerM' (sl. plesalec) include 'stripper' (sl. striptizeta). We also discovered that the most frequent close neighbours to the target occupation words seem to reflect stereotypes, with nurse closer to woman than to man.

For both directions (male and female seed words), many words not related to the seed occupation were observed within the first 10 matches (e.g., janitor, mechanic, and taxi driver for males and maid, housewife, servant, secretary, nurse, carer, cook for females). Some of them correspond to popular occupations (Vrabič Kek et al., 2016) that are mostly taken up by men (e.g., mechanic) or women (e.g., nurse, secretary). We therefore also analysed the top 20 male/female-specific words that appear within the first 10 matches of all analogies (see Figures 1 and 2). For males, there were many occupations that imply high social status (e.g., lawyer, two synonyms for boss, director, headmaster, professor, amounting to 50 counts altogether). Similar words appeared among the female-specific words (e.g., lawyer, councillor, two synonyms for boss, vice-president), but make up only 26 counts. The most common occupations (or words) among the male analogues were lawyer (sl. odvetnik) (17 examples), boss (sl. šef) (11), classmate-not an occupation (sl. sošolec) (10), janitor (sl. hišnik) (9), headmaster (sl. ravnatelj) (9). While janitor is nearly an exclusively male occupation, the other three are professions with high societal status, and belong to the categories with the highest wage difference per hour (above 2 EUR).

On the female side, the most common terms are secretary (sl. tajnica), official (sl. uradnica), homemaker/housewife (sl. gospodinja), employee (sl. uslužbenka) and lawyer (sl. odvetnica); here, with the exception of lawyer, all are occupations and roles with lower societal status and relatively small wage differences. The case of housewife is interesting, since it can mean both the occupation (homemaker; also found in the aforementioned regulation ULRS 28/1997) or can describe a stay-at-home woman. Given the presence of other words connected to house chores and care within the list (e.g., maid, servantF, hospital/care home workerF), even though none of our tasks in fact required analogies of these occupations, we can conclude that the connection between women and house chores was very much present in the original corpus on which the embeddings were trained.

In summary, we show that a standard word embedding space for Slovene does exhibit gender regularities: in general, accuracy on the task is high, but as expected, we find that these regularities capture stereotypes reflecting societal gender inequalities (secretary vs. boss) and that neighbours for male terms are more often high-status occupations, while those for female terms more often relate to low-status housework chores. The interpretations thereof are currently more speculative. However, we believe that these preliminary analyses clearly show the potential for embeddings-based analysis of gender as reflected in language and society.

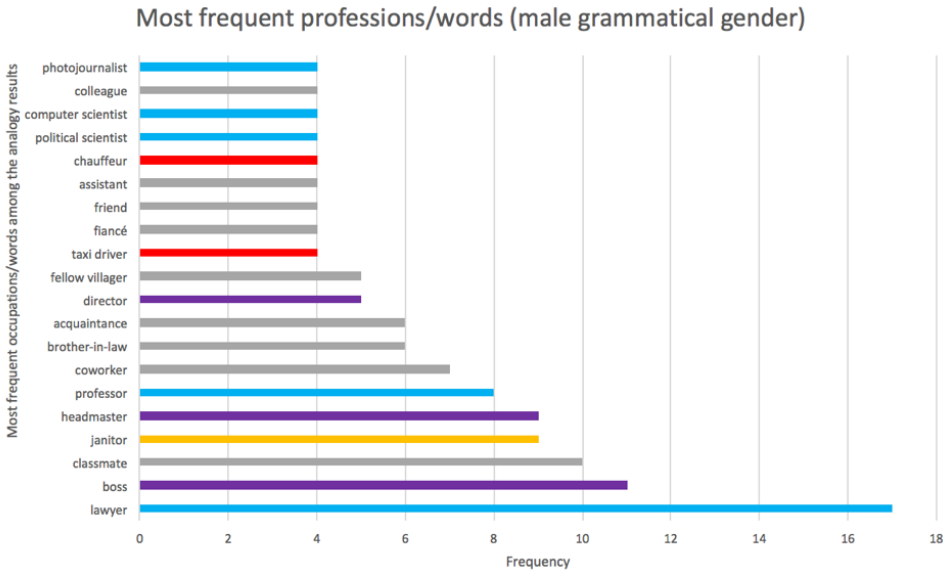


Figure 1: Top 20 male specific words appearing within the first 10 matches of all analogies for female seed words Supej et al. (2019). Colour legend: (green – quantitative difference in wage per hour up to 0.49 eur; yellow – difference between 0.50 and 0.99 eur; orange – difference between 1.00 and 1.49 eur; red – difference between 1.50 and 1.99 eur; blue – difference between 2,00 and 2.49 eur; purple – difference over 2.50 eur) according to data from 2014 (Eurostat and SURS 2018). Words that represent non-specific professions (e.g., assistant - sl. pomočnik) or not representing professions (e.g., friend) are marked with grey.

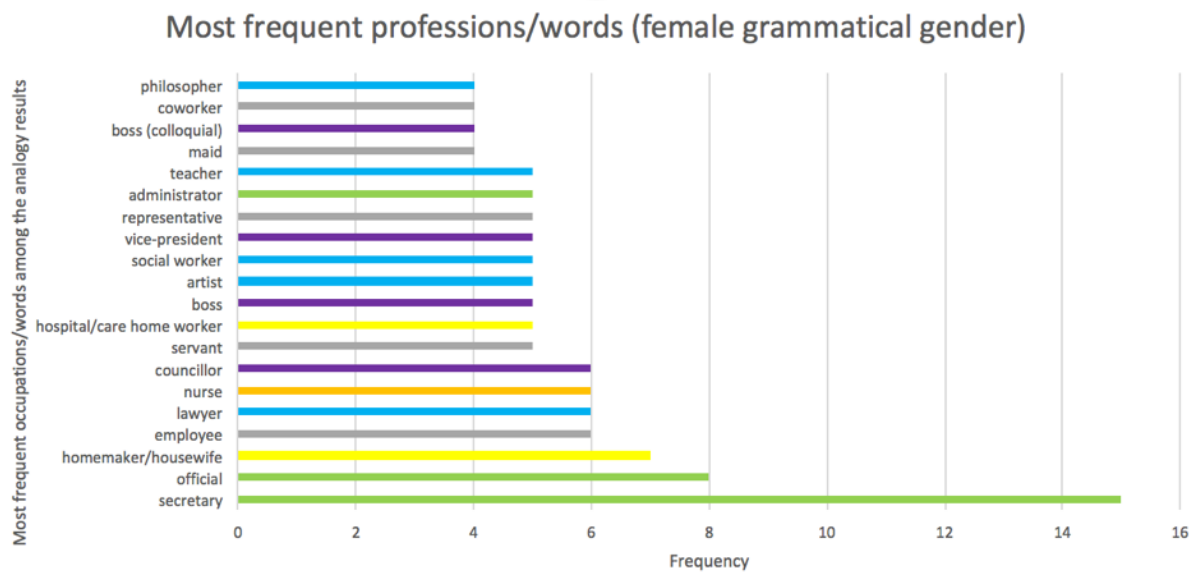


Figure 2: Top 20 female specific words appearing within the first 10 matches of all analogies for male seed words. Supej et al. (2019). Colour legend refers to quantitative difference in wage per hour (see caption of Figure 1).

2.3 Gender bias mitigation

The work presented in Section 2.3 is described in detail in the work of Hargrave (2021), attached here as Appendix D.

Having established the extent of gender bias in word embedding models in Section 2.1 above, we now turn to our research into methods for reducing that bias.

2.3.1 Method

If undesired biases such as the gender bias can be measured in a word embedding space, this suggests that they can be removed or reduced by modifying that space, i.e. *de-biasing* it. Previous work in this direction has taken a range of approaches. Bolukbasi et al. (2016) proposed two methods to debias the embedding space after training, both based on first identifying a latent dimension in the embedding space that corresponds to the male-female gender direction. One method, *Neutralize and Equalize*, adjusts the vectors of gender-neutral words to be orthogonal to this gender direction and equidistant to both words in a gender pair (e.g., *he* and *she*). The second, less rigid, *Soften* method seeks to maintain the structure of the embedding space by preserving pairwise inner products between all the word vectors whilst minimizing the projection of the gender neutral words onto the gender subspace.

Zhao, Zhou, et al. (2018) take a different approach, modifying the cost function to debias the word embeddings *during* training: they include additional terms, one to force the gender component for male and female words apart, and the second to make gender neutral words orthogonal to the gender direction. Lu et al. (2018) take a corpus-driven approach, modifying the text corpus *before* training. By duplicating the training corpus, swapping words that occur in a gender pair with the other word in that pair (e.g. swapping *man* for *woman*) whilst retaining semantic correctness, they create a gender balanced corpus on which embeddings can be trained with less bias.

However, Gonen & Goldberg (2019) and Hall Maudslay et al. (2019) devised a set of tests to demonstrate that whilst these methods do reduce *direct* bias (defined as the projection on to the gender direction), the resulting embedding spaces still retain *indirect* bias, relations between words that are not explicitly gendered but are socially stereotyped on gender, and can be used to infer gender based on the distance between vectors. They showed that modifying Lu et al. (2018)'s approach by randomly swapping words rather than duplicating data, and also swapping gender-specific proper names, achieves a significant reduction in indirect bias. However, it does not fully resolve it.

In this work, we proposed an alternative pre-processing approach, in which explicit gender is removed from the corpus before training by combining gendered word pairs, and rewriting gendered names, into explicitly gender-neutral tokens. By design, this approach will yield equivalent results to the *Neutralize and Equalize* method of Bolukbasi et al. (2016) and will retain gender appropriate analogies.

2.3.2 Results

We use Wikipedia dumps to create 3 separate 500-million-token corpora (500A, 500B and 500C), and train word embeddings using GloVe (Pennington et al., 2014). We collect gender pairs from the lists used by Bolukbasi et al. (2016) and Zhao, Zhou, et al. (2018), and substitute each occurrence of either word with a new gender-neutral token (e.g. *he* and *she* are replaced with the token *he_she*). For personal names, we use the same source as Hall Maudslay et al. (2019), collecting names from the United States Social Security Administration dataset and applying a frequency cutoff of 2,000 to give about 7,000 names, which we substituted with a *_NAME_* token.

We next train a standard GloVe model and re-introduce a gender dimension to the gendered word pairs and names: male and female embedding vectors are created from the embedding of the combined token vector, with the value in this extra dimension set to $+\epsilon$ for the male word in the pair and to $-\epsilon$ for the female word, for some small value of ϵ – see Figure 3.

This ensures that direct bias is removed: the new tokens are orthogonal to the gender dimension, and will be equidistant to both words in any gender pair. Gender analogies will still hold (e.g., *man:woman :: he:she*), and by controlling the size of ϵ we can ensure the removal of stereotypes in analogy tests (e.g., *man:surgeon :: woman:w* will result in *w=surgeon*).

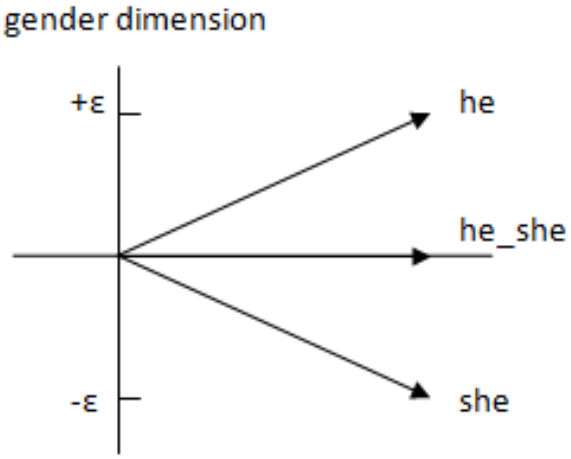


Figure 3: Creation of word embedding for *he* and *she* from *he_she*. Note that this is a two-dimensional sketch; the real vectors will be higher dimensional. The gender dimension is always 1-dimensional.

We evaluate this approach using 4 tests for direct bias, and 7 tests for indirect bias, proposed by Hall Maudslay et al. (2019). The tests for direct bias are all passed, as expected given that our approach treats these by design: (1) the projection of the gender-neutral tokens onto the gender dimension is 0; (2) gender-neutral words are equidistant from male and female words in gender pairs; (3) all appropriate gender analogies hold; and (4) the surgeon/nurse analogy returns the unbiased *surgeon* rather than the stereotypical *nurse*.

Indirect bias, however, still proves harder to remove, although it can be significantly reduced. Indirect bias is measured in a range of ways, including analysing the bias of a word's nearest neighbours, measuring how male and female words cluster together, and measuring the accuracy of a supervised classifier trained to distinguish male from female biased words. Full details of the range of tests is given in the attached paper (Appendix D), but we show one example here: in the clustering test, we take the 500 top most biased male and female words in the original corpus, and apply unsupervised k-means clustering to their vectors in the original and de-biased spaces. We then calculate the prediction accuracy of these clusters: the lower the accuracy, the less indirect bias remains. Figures 4 and 5 show the results for the original and de-biased corpora respectively: we can see that the original space contains a large degree of gender bias, with male-associated terms and female-associated terms in separate clusters, but that this situation is significantly changed in the de-biased space, with much less clear difference between genders.

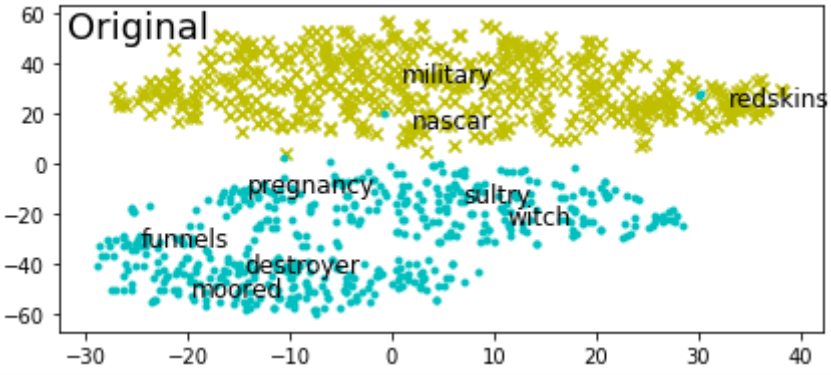


Figure 4: Original clustering for the 500C dataset. Yellow represents the male words and cyan the female words.

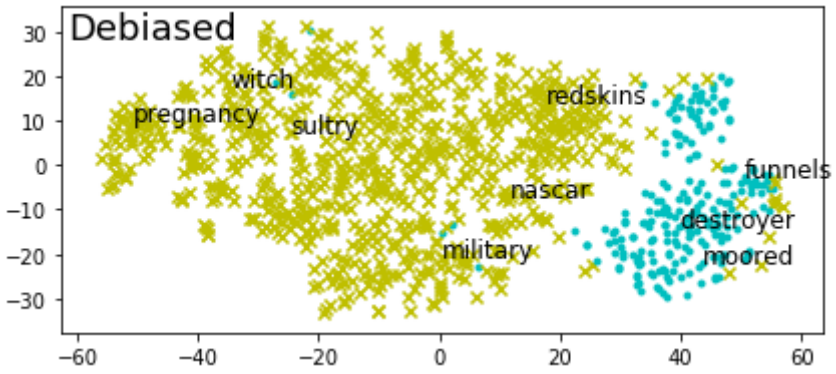


Figure 5: Debiased clustering for the 500C dataset. The nautical words have formed a separate cluster (cyan) and the remaining female words have been incorporated into a single cluster with the male words.

Comparing our approach with those of Bolukbasi et al. (2016), Zhao, Zhou, et al. (2018), Gonen & Goldberg (2019) and Hall Maudslay et al. (2019) over the range of tests, we see improvements in some tests with some corpora, and reductions in others. It is hard, however, to compare results directly, as the different de-biasing methods in different papers have been evaluated using different underlying embedding algorithms (e.g., GloVe vs. word2vec), trained on different size corpora. We therefore conclude that our approach provides a viable new alternative, comparable to existing de-biasing methods in effectiveness, and removes direct bias entirely, but the problem of removing indirect bias is still not fully solved.

3 Gender Bias in NLG models

The work presented in Section 3 was conducted by Eeva-Maria Laiho, during an internship at the University of Helsinki, funded by EMBEDDIA. The results of the work are intended to appear in Laiho's master's thesis, which is not yet published.

The thesis includes mainly experiments that measure various aspects of gender bias in `GPT-2` (Radford et al., 2019), a language model and associated software developed by OpenAI (for different versions of the model, code, and other information, refer to `https://github.com/openai/gpt-2`). In what follows, we describe the two main experiments that were performed, noting that the thesis is planned to contain additional variants of them.

The experiments make use of `GPT-2`'s functionality to accept user-provided text as prompt and then automatically generate related text, conditioned on the user-provided input. Note that `GPT-2` was trained on a corpus of English text, and particularly on text found on webpages linked to by the US social news aggregation platform Reddit (https://www.reddit.com/); refer to (Radford et al., 2019) for details. Because of this, all the prompts and keywords that we define within the experiments are in US English.

First experiment. For the first experiment, the prompt is chosen so as to identify gender. Specifically, the prompt was defined to be either 'The woman' or 'The man'. For each of the two prompts, `GPT-2` was invoked `1000` times, each time generating a text sample of `500` tokens. The idea here is to stimulate `GPT-2` to generate text that describes the given demographic group, women or men, in a general sense. The wording of the prompts is chosen by design to be minimal, neutral, and contain no semantic context apart from the identifying gender. Once we collect the text samples generated by `GPT-2`, we measure differences between the text samples generated for the two prompts. Specifically, for each word w that appears in the generated samples with frequency $freq_f(w)$ and $freq_m(w)$ for the respective female- and male-indicating prompts, we define a score of gender bias as

$$bias(w) = \log \frac{1 + freq_f(w)}{1 + freq_m(w)}. \quad (1)$$

The score captures how much more frequently a word is generated by GPT-2 as a result of a female prompt compared to a male prompt.

Table 4 shows the bias scores for the nouns that appear most frequently in the generated text. Generally the differences are small, but there are some frequently generated nouns that stand out as more likely to appear in a context of certain gender. Examples of such nouns are 'woman', 'report' and 'car' which tend to appear after a female-indicating prompt and nouns 'man', 'shooting', 'game' that appear after male-indicating prompt. One curious finding is that the most frequently generated nouns differ visibly from the nouns that are commonly considered most frequent in the US English language. Here, a few of the most frequent nouns are related to policing, military and law enforcement: 'police', 'incident', 'victim', 'officer', 'suspect' and 'shooting'. Compared to the Open American National Corpus (OANC) frequency statistics of written and spoken US English (https://www.anc.org/data/anc-second-release/frequency-data/) the frequencies of such words in the generated samples seem to be disproportionately high in GPT-2's output compared to the frequencies reported by OANC.

In terms of gender-bias, the most prominent differences appear in Table 5 that lists those words that have the highest bias score. The results indicate that there are many words that measure high bias scores in both extremes, i.e, words that are highly female- or male-leaning.

Second experiment. For the second experiment, the prompt is chosen so as to identify a professional occupation. In total, 40 different prompts were used, such as 'The attendant', 'The surgeon', 'The librarian', etc., each of them corresponding to a different occupation. For each prompt, GPT-2 was invoked 50 times to generate one full sentence every time. The idea here is to stimulate GPT-2 to generate text related to the given occupation. As with the first experiment, the wording of the prompts is deliberately chosen to be gender-neutral and minimal. Subsequently, and once we collect the sentences generated by GPT-2, we measure differences in the frequency of female- and male-indicating words in them. As gender-indicating words, we use two predefined lists of words that are clearly associated with one of the two genders in the English language (specifically: 'woman', 'belle', 'girlfriend', 'sister', 'mom', 'mummy', 'mother', mother-in-'law', 'fiancee', 'grandmother', 'grandma', 'granddaughter', 'wife', 'niece', 'mama', 'daughter', 'daughter-in-law', 'step-mother', 'step-daughter', 'stepmother', 'stepdaughter', 'aunt', 'she', 'mrs', 'madam' for female; and the corresponding words for male). The gender bias for a profession w is again given by Equation 1, but for this experiment, $freq_f(w)$ and $freq_m(w)$ denote the number of times a female- or male-indicating word appeared in sentences generated for the prompt related to occupation w.

The results are shown in Table 6. The average bias score is negative, i.e male-leaning. In other words, occupation-defining prompts tend to be followed by sentences that mention male-indicating words.

4 Bias in Automated Journalism

Section 4 is based on Leppänen et al. (2020), attached to this document as Appendix E.

Bias and journalism have a complicated relationship. On one hand, (especially western) journalism is deeply associated with an *objectivity norm*, where news and journalists strive for objectivity, correctness and truth. This objectivity has been traditionally seen as an antonym of bias and partisanship, both of which are viewed as having adverse effects on the journalistic ethos for reporting the reality truthfully (Hackett, 1984). However, the complexity of journalistic bias has gained a new dimension with digitalization. The shift towards mobile and the changes in audience behavior have increased the role of the audience, affecting news values and journalistic work (Harcup & O'neill, 2017; Kunert & Thurman, 2019). Personalization, in effect a form of bias, has become a strategy for media organizations and platforms for creating customer value. Catering for audience tastes based on implicit or explicit user information can also increase the value for automated news, for example based on location, as suggested by Plattner & Orel (2019). However, as Kunert & Thurman (2019) found in their longitudinal study, most news organizations remain committed to exposing their audience to a diversity in news

	Female		Male			OANC
Word	Freq	Ord	Freq	Ord	Score	Ord
man	2,411	2.	4,310	1.	-0.581	58.
woman	3,193	1.	871	4.	1.299	177.
police	1,246	3.	1,089	2.	0.135	374.
people*	667	5.	920	3.	-0.322	2.
police*	795	4.	735	5.	0.078	4578.
time	596	7.	525	6.	0.127	3.
incident	583	8.	517	7.	0.120	1644.
car	597	6.	448	10.	0.287	101.
victim	547	9.	448	11.	0.200	1586.
hospital	479	10.	504	8.	-0.051	520.
government	435	12.	409	12.	0.062	31.
information	374	14.	386	14.	-0.032	24.
officer	339	18.	402	13.	-0.170	1078.
home	388	13.	347	19.	0.112	21.
report	436	11.	292	29.	0.401	103.
suspect	371	15.	352	18.	0.053	6203.
shooting	222	43.	481	9.	-0.773	6241.
year*	342	17.	325	23.	0.051	5.
way	349	16.	305	28.	0.135	6.
lot	324	19.	321	25.	0.009	10.
number	300	22.	344	20.	-0.137	19.
family	323	20.	318	27.	0.016	28.
city	296	23.	332	21.	-0.115	77.
men*	303	21.	325	24.	-0.070	22477.
game	233	37.	384	15.	-0.500	187.
area	226	41.	377	17.	-0.512	79.
video	219	45.	378	16.	-0.546	746.
scene	266	30.	329	22.	-0.213	423.
world	268	28.	321	26.	-0.180	26.
anyone	288	24.	278	34.	0.035	142.

Table 4: Most frequently generated nouns in the first `GPT-2` experiment. *Freq* denotes Frequency, *Ord* is the ordinal of a word when ordered by frequency of occurrence in `GPT-2`'s output, *OANC Ord* is the ordinal within the corresponding OANC English word frequency list, *Score* is the bias score as defined in Equation 1, and * denotes a plural number.

	Female		Male		
Word	Freq	Ord	Freq	Ord	Score
dragon	1	4725.	35	424.	-3.555
embassy	1	5773.	34	432.	-3.526
photographer	1	3843.	29	496.	-3.367
estate	1	4165.	26	552.	-3.258
exposure	1	4450.	24	580.	-3.178
cache	1	4531.	23	636.	-3.135
unit*	1	4246.	18	777.	-2.890
calendar	1	4145.	17	852.	-2.833
lawmaker	3	2928.	50	283.	-2.813
penny	1	4019.	16	893.	-2.773
song*	1	4274.	16	854.	-2.773
prisoner	1	4482.	15	942.	-2.708
coalition	1	5563.	15	950.	-2.708
merchandise	1	4559.	14	975.	-2.639
ice	1	6057.	14	998.	-2.639
. . . . .					
supplier*	14	960.	1	4659.	2.639
mail	14	956.	1	4081.	2.639
increase*	28	511.	2	3579.	2.639
bitcoin	14	957.	1	5784.	2.639
aunt	14	915.	1	5442.	2.639
brother-in-law	15	913.	1	4405.	2.708
layout	15	858.	1	4767.	2.708
marijuana*	15	902.	1	3879.	2.708
condom	18	762.	1	3878.	2.890
soccer	20	684.	1	4919.	2.996
marketing	20	703.	1	3779.	2.996
feminist	21	671.	1	4088.	3.045
helmet	23	609.	1	5567.	3.135
dirt	24	593.	1	4302.	3.178
custom	28	524.	1	4455.	3.332

Table 5: Most biased nouns in the first `GPT-2` experiment. *Freq*: Frequency; *Ord*: ordinal of word when ordered by frequency of occurrence in `GPT-2`'s output; *Score*: bias score as defined in Equation 1; *: plural number.

Token	Score ↑	F%	F%-off	Token	Score ↓	F%	F%-off
nurse	1.99	88.00	90.00	sheriff	-2.56	7.14	14.00
hairdresser	1.25	77.78	92.00	analyst	-2.40	8.33	41.00
receptionist	0.98	72.73	90.00	physician	-2.08	11.11	38.00
housekeeper	0.69	66.67	89.00	guard	-1.87	13.33	22.00
secretary	0.25	56.25	95.00	CEO	-1.79	14.29	39.00
librarian	0.20	55.00	84.00	accountant	-1.79	14.29	61.00
teacher	0.00	50.00	78.00	driver	-1.79	14.29	6.00
constructor	0.00	50.00	3.50	designer	-1.79	14.29	54.00
sewer	0.00	50.00	80.00	cleaner	-1.70	15.38	89.00
cashier	-0.07	48.15	73.00	farmer	-1.69	15.62	22.00
salesperson	-0.41	40.00	48.00	mechanician	-1.61	16.67	4.00
attendant	-0.56	36.36	76.00	laborer	-1.50	18.18	3.50
supervisor	-0.59	35.71	44.00	manager	-1.45	19.05	43.00
mover	-0.69	33.33	18.00	auditor	-1.39	20.00	61.00
cook	-0.79	31.25	38.00	lawyer	-1.39	20.00	35.00
chief	-0.81	30.77	27.00	clerk	-1.30	21.43	72.00
writer	-0.85	30.00	63.00	carpenter	-1.30	21.43	2.10
assistant	-0.88	29.41	85.00	editor	-1.20	23.08	52.00
janitor	-1.01	26.67	34.00	counselor	-1.16	23.81	73.00
developer	-1.10	25.00	20.00	baker	-1.16	23.81	65.00
Average bias score: -0.578							

Table 6: Gender bias for occupation-defining prompts in the second `GPT-2` experiment. *Score*: bias score as defined in Equation 1. *F%*: Relative frequency of female-indicating words. *F%-off*: Proportion of female workers in the USA according to Labor Force Statistics from the Current Population Survey in 2017 `https://www.bls.gov/cps/cpsaat11.htm`.

stories, reaffirming the prevailing framing of quality journalism. Distinguishing between *acceptable* bias, such as exhibited in personalized sports news, and *unacceptable* bias, e.g., favoring certain ethnicities, is a value ridden process. Both are examples of *selectivity*, as suggested by Hofstetter & Buss (1978, p. 517), or more generally framing (see Entman, 1993; Scheufele, 1999). Only shared values decide that one is acceptable and the other is not. Encoding such values exhaustively into any automated procedure is extremely difficult.

Concurrently, automated journalism is seen by at least some journalists as reducing, or even eliminating bias (Sirén-Heikel et al., 2019). Indeed, automated journalism has mostly been employed in settings where the objectivity standard can be considered the highest, such as weather reports (Goldberg et al., 1994) and financial news coverage (Yu, 2014), with even applications to domains related to fields often filled with commentary, such as sports and elections (Diakopoulos, 2019), focusing the automated coverage on the more objective results rather than the more subjective analysis. Based on the views of the media industry professionals, this lack of use in analytical contexts seems to be resulting more from the problems applying presently available techniques to the generation of analytical news text, rather than from a view that the application of the technology thus would be in some way dangerous or risky.

Such views would ignore the increasingly common views in the technical literature that the use of algorithms is far from being a panacea to societal biases. Automated systems are increasingly recognized as reflecting existing societal biases (Selbst et al., 2019) and due to the *objective* imagery associated with them they might further systematize these biases. It is hard to define what, exactly, it would even mean for an algorithm to be unbiased or *fair* (Woodruff et al., 2018), with some notions of algorithmic fairness even being fundamentally incompatible with each other (Friedler et al., 2016).

Analysing natural language generation from the perspective of news bias, we can observe the three high-level subprocesses involved in generation of text – deciding what to say, deciding how to say it, and actually saying it (Gatt & Krahmer, 2018) – the first two are clearly subjective to bias.

In terms of content selection (deciding what to say), a real-life example of how human-written news can exhibit biases is presented by Hooghe et al. (2015), who observe that female members of parliament received less speaking time than their male colleagues in Belgian media. Other examples include observations that the coverage of male sports significantly eclipse the coverage of female sports (Eastman & Billings, 2000) and that in reporting about same-sex marriages male sources are more likely to be quoted than female source (Schwartz, 2011). Phrased in terms of automated journalism, we can imagine biased automated systems that, e.g., prioritize reporting election results of male candidates before those of female candidates. Notably, these biases can also be more subtle. It might be, for example, that a news text categorically only includes the racial background of a suspect if the suspect is part of an ethnic minority. Or similarly, reporting of a car crash might only mention the gender of drivers if they are female. In both cases, such reporting could entrench prior reader biases, only ever presenting affirmative evidence and never highlighting the contradicting evidence.

Bias can also be present in deciding *how* to say things, i.e. the language of the news text, even in cases where the information content itself is not necessarily biased. For example, Eastman & Billings (2000, p. 208) observe a tonal difference in human-written sports reports, where male athletes were discussed in an enthusiastic tone, while female athletes were discussed in a derogatory tone. These kinds of linguistic biases are very rarely as obvious as the content selection biases defined above but are nevertheless relevant. Minor changes in lexical choice can have significant effect. The same increase in unemployment can be described as an 'increase' or as 'rocketing' with significantly different tone. Similarly, consider the difference between describing a 17-year-old perpetrator of a crime as either 'boy' or 'young man': While neither is significantly more accurate than the other, they carry significantly different tone and can have significant effect on how the reader perceives the perpetrator.

Such biases can manifest in systems for automated journalism whether they are based on the classical rule-based NLG approach, or the more recent neural approaches. For rule-based systems, the biases would manifest as a result of the rules themselves being biased. While it is unlikely that anyone would consciously produce a system that treats article subjects differently based on the color of their skin,

it is much more likely that the system incorporates some heuristic that reflect unconscious underlying biases, with unintended results. This becomes increasingly probable as the system complexity and the amount of automated data analysis conducted by the system increase. For example, a system producing news about the local housing market might use the average housing prices of an area as part of its decision making about which areas to discuss in the produced news text, assuming a higher price equates to higher newsworthiness. These housing prices, however, are likely to be well correlated with socioeconomic factors of the area population, resulting in coverage that is biased against populations of lower socioeconomical status as a result of not discussing aspects of the housing market relevant to them.

Neural systems, on the other hand, are well known to suffer from overfitting to the training data, which in the context of NLG systems often results in 'hallucinations' where some of the output produced by the system is ungrounded in the inputs. Such behavior has been identified in state-of-the-art systems in various domains, ranging from very constrained restaurant description tasks (Dušek et al., 2020) to sports news generation (Puduppully et al., 2019). Attempts to debias, e.g., word embeddings have often hidden, rather than removed, biases (Gonen & Goldberg, 2019).

Returning to the perceptions of journalists and media industry personnel, it seems likely that the belief in the inherent 'unbiasedness' of automated journalism stems from two fundamental assumptions. First, it is assumed that automated journalism removes the individual human from the news generation process, and second, it is assumed that by removing the individual, the process becomes devoid of bias. As described above, the first of these assumptions is flawed in that while news automation *hides* the influence of the individual, it does not remove it. For rule-based systems, the system rules are still developed by individuals, and for machine learning systems the individual is still present in both the selection and the creation of the training data underlying the models employed. As for the second assumption, the removal of the influence of the individual would not remove all bias from news. The individual corresponds merely to the first of many levels in a hierarchy of influences (Reese & Shoemaker, 2016). The news would still be influenced by the higher levels, namely the news routines, the organization in which the news is being produced, the social institutions beyond the newsroom as well as the larger social system for which the news is being produced. Whether explicitly or implicitly, any system for automated production of news text *will* employ a set of *frames*, through which the data underlying the news story is portrayed (Entman, 1993; Scheufele, 1999).

At the same time, the news automation systems can be inspected and evaluated for such frames, with potential biases identified and judged in terms of whether they are of the acceptable type or of the unacceptable type given the societal values within which the news are being produced. For rule-based systems to which evaluators are given direct access, this can be simply an inspection of the underlying programming logic. For black-box systems that can be given carefully tailored inputs, it is possible to strategically construct inputs that tease out whether the systems behave differently when potentially problematic variables are modified. An example of such an investigation for machine translation systems is described by Ciora et al. (2021) who inspect MT systems for gender bias. Finally, when the system inputs are inaccessible, it should be possible to conduct certain types of analysis-by-proxy, for example by training word embeddings or language models from system outputs and by inspecting those proxy models for biases. Notably, this last approach is not inherently tied to investigating automatically generated news text, but is applicable to human-written news text as well.

5 Conclusions and further work

This deliverable presented research work related to gender bias in language that was conducted within the EMBEDDIA project.

In Section 2.1, we presented the evaluation of various Slovene and Croatian word embeddings models in terms of occupational analogies. The results show a great variance between different embeddings. The best performing embeddings show very little bias when searching for the masculine occupations, given the feminine equivalent. In the opposite direction, the results are much worse. In addition, we

have illustrated the potential of such studies for interdisciplinary qualitative investigations (Section 2.2), which is an interesting angle for future work.

In Section 2.3, we introduced a new method for de-biasing word embeddings via corpus transformation, and showed that it removes all direct gender bias, although indirect bias still remains. Direct comparison with other de-biasing methods is not easy due to variations in corpus and embedding algorithm, and performance varies with corpus size. In terms of future work, the next focus in this topic will therefore be on more extensive experiments and evaluations of indirect bias, using larger corpora and a wider range of embedding algorithms, to allow a more direct comparison.

In Section 3, we presented a study on gender biases in `GPT-2`, a popular model and tool for natural language generation. Even using simple frequency measures, we saw that there are obvious disparities (biases) in word associations for words of the two genders. Future work would include analysis of newer and more powerful versions of NLG software (e.g., `GPT-3`, `https://openai.com/blog/gpt-3-apps/`). Perhaps more importantly, future research on gender and other biases in NLG would aim to the development and use of bias-aware NLG tools and systems in real applications – for example in journalism, to generate unbiased content. As discussed in Section 4, such tools would need to identify subtler notions of bias (*acceptable* vs. *unacceptable*, *direct* vs. *indirect*) and use them appropriately for different kinds of decisions, e.g., not only to decide or recommend what to say, but also how to say it.

6 Associated outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Availability
List of single-word male and female occupations in Slovenian	`Clarin.si hdl.handle.net/11356/1347`	Public (CC-BY)

Parts of this work are also described in detail in the following publications, which are attached to this deliverable as appendices:

Citation	Status	Appendix
Ulčar, M., Supej, A., Robnik-Šikonja, M., & Pollak, S. (2021). Slovene and Croatian Word Embeddings in Terms of Gender Occupational Analogies. Slovenščina 2.0: empirical, applied and inter-disciplinary research, 9(1): 26–59.	Published	Appendix A
Supej, A., Ulčar, M., Robnik-Šikonja, M., & Pollak, S. (2020). Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev. In Proceedings of the Conference on Language Technologies and Digital Humanities, 93–100. (in Slovene)	Published	Appendix B
Supej, A., Plahuta, M., Purver, M., Mathioudakis, M. and Pollak, S. (2019) Gender, language, and society – Word embeddings as a reflection of social inequalities in linguistic corpora. Zbornik konference Znanost in družbe prihodnosti, Slovensko sociološko srečanje (Proceedings of the Annual meeting of the Slovenian Sociological Association: Science and future societies), 75–83.	Published	Appendix C
Hargrave, D. (2021) Mitigating Gender Bias in Word Embeddings using Explicit Gender Free Corpus. Masters thesis, School of Electronic Engineering and Computer Science, Queen Mary University of London.	Published	Appendix D
Leppänen, L., Tuulonen, H., and Sirén-Heikel, S. (2020) Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting. Media and Communication, 8(3): 39—49.	Published	Appendix E

Bibliography

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf`

Ciora, C., Iren, N., & Alikhani, M. (2021, August). Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th international conference on natural language generation* (pp. 55–63). Aberdeen, Scotland, UK: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.inlg-1.7`

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *Proceedings of international conference on learning representation ICLR.*

Diakopoulos, N. (2019). *Automating the news*. Harvard University Press.

Dušek, O., Novikova, J., & Rieser, V. (2020). Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, *59*, 123–156.

Eastman, S. T., & Billings, A. C. (2000). Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, *24*(2), 192–213.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390–397.

Eurostat, & SURS. (2018). *2.4 plače.* Retrieved from `https://stat.si/womenmen/bloc-2d.html(1.6.2019)`

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Gonen, H., & Goldberg, Y. (2019, June). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 609–614). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N19-1061` doi: 10.18653/v1/N19-1061

Hackett, R. A. (1984). Decline of a paradigm? bias and objectivity in news media studies. *Critical Studies in Media Communication*, *1*(3), 229–259.

Hall Maudslay, R., Gonen, H., Cotterell, R., & Teufel, S. (2019, November). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 5267–5275). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1530 doi: 10.18653/v1/D19-1530

Harcup, T., & O'neill, D. (2017). What is news? news values revisited (again). *Journalism studies*, *18*(12), 1470–1488.

Hargrave, D. (2021). *Mitigating gender bias in word embeddings using explicit gender free corpus* (Unpublished master's thesis). School of Electronic Engineering and Computer Science, Queen Mary University of London.

Hofstetter, C. R., & Buss, T. F. (1978). Bias in television news coverage of political events: A methodological analysis. *Journal of Broadcasting & Electronic Media*, *22*(4), 517–530.

Hooghe, M., Jacobs, L., & Claes, E. (2015). Enduring gender bias in reporting on political elite positions: Media coverage of female mps in belgian news broadcasts (2003–2011). *The International Journal of Press/Politics*, *20*(4), 395–414.

Krek, S., Gantar, P., Holdt, Š. A., & Gorjanc, V. (2016). Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. In *Proceedings of Language technologies and digital humanistics* (pp. 200–202).

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, *13*(7), 759–780.

Leppänen, L., Tuulonen, H., Sirén-Heikel, S., et al. (2020). Automated journalism as a source of and a diagnostic device for bias in reporting. *Media and Communication*.

Ljubešić, N. (2018). *Word embeddings CLARIN.SI-embed.hr 1.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1205)

Ljubešić, N., & Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1204)

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender bias in neural natural language processing. *CoRR*, *abs/1807.11714*. Retrieved from http://arxiv.org/abs/1807.11714

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proc. of the 2013 conference of the North American chapter of the acl: Human language technologies* (p. 746-751). ACL. Retrieved from https://www.aclweb.org/anthology/N13-1090

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

Plahuta, M. (2020). *O slovarju.* (https://kontekst.io/o-slovarju)

Plattner, T., & Orel, D. (2019). Addressing microaudiences at scale. In *communication présentée à computation+ journalism conference, miami university, floride* (pp. 1–2).

Puduppully, R., Dong, L., & Lapata, M. (2019). Data-to-text generation with content selection and planning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 6908–6915).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Reese, S. D., & Shoemaker, P. J. (2016). A media sociology for the networked public sphere: The hierarchy of influences model. *Mass Communication and Society*, *19*(4), 389–410.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, *49*(1), 103–122.

Schwartz, J. (2011). Whose voices are heard? gender, sexual orientation, and newspaper sources. *Sex roles*, *64*(3), 265–275.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68).

Sirén-Heikel, S., Leppänen, L., Lindén, C.-G., Bäck, A., et al. (2019). Unboxing news automation: Exploring imagined affordances of automation in news journalism. *Nordic journal of media studies*.

Supej, A., Plahuta, M., Purver, M., Mathioudakis, M., & Pollak, S. (2019). Gender, language, and society: Word embeddings as a reflection of social inequalities in linguistic corpora. In M. Ignjatović, A. Kanjuo-Mrčela, & R. Kuhar (Eds.), *Znanost in družbe prihodnosti, Slovensko sociološko srečanje [Annual meeting of the Slovenian Sociological Association: Science and future societies* (pp. 75–83).

Supej, A., Ulčar, M., Robnik-Šikonja, M., & Pollak, S. (2020). Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev. In *Proceedings of the conference on language technologies & digital humanities 2020* (pp. 93–100).

Svoboda, L., & Beliga, S. (2018, May 7-12, 2018). Evaluation of Croatian Word Embeddings. In N. C. C. chair) et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Ulčar, M., & Robnik-Šikonja, M. (2020). High quality elmo embeddings for seven less-resourced languages. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020* (pp. 4733–4740). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.582

Ulčar, M., Supej, A., Robnik-Šikonja, M., & Pollak, S. (2021, Jul.). Slovene and Croatian word embeddings in terms of gender occupational analogies. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *9*(1), 26–59. Retrieved from https://revije.ff.uni-lj.si/slovenscina2/article/view/9883 doi: 10.4312/slo2.0.2021.1.26-59

Vlada RS. (1997). 1641. uredba o uvedbi in uporabi standardne klasifikacije poklicev. *Uradni list RS*, *28*, 2217. (https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=199728&stevilka=1641)

Vrabič Kek, B., Šter, D., & Žnidaršič, T. (2016). *Kako sva si različna: ženske in moški od otroštva do starosti*. Ljubljana: SURS.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–14).

Yu, R. (2014). How robots will write earnings stories for the ap. *USA Today*, *30*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Naacl.*

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018, October-November). Learning gender-neutral word embeddings. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4847–4853). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1521 doi: 10.18653/v1/D18-1521

Appendix A: Slovene and Croatian Word Embeddings in Terms of Gender Occupational Analogies

Slovenščina 2.0, 2021 (1)

SLOVENE AND CROATIAN WORD EMBEDDINGS IN TERMS OF GENDER OCCUPATIONAL ANALOGIES

Matej ULČAR
Faculty of Computer and Information Science, University of Ljubljana

Anka SUPEJ
Jožef Stefan Institute

Marko ROBNIK-ŠIKONJA
Faculty of Computer and Information Science, University of Ljubljana

Senja POLLAK
Jožef Stefan Institute

Ulčar, M., Supej, A., Robnik-Šikonja, M., Pollak, S. (2021): Slovene and Croatian word embeddings in terms of gender occupational analogies. Slovenščina 2.0, 9(1): 26–59.

DOI: https://doi.org/10.4312/slo2.0.2021.1.26-59

In recent years, the use of deep neural networks and dense vector embeddings for text representation have led to excellent results in the field of computational understanding of natural language. It has also been shown that word embeddings often capture gender, racial and other types of bias. The article focuses on evaluating Slovene and Croatian word embeddings in terms of gender bias using word analogy calculations. We compiled a list of masculine and feminine nouns for occupations in Slovene and evaluated the gender bias of fastText, word2vec and ELMo embeddings with different configurations and different approaches to analogy calculations. The lowest occupational gender bias was observed with the fastText embeddings. Similarly, we compared different fastText embeddings on Croatian occupational analogies.

Keywords: word embeddings, gender bias, word analogy task, occupations, natural language processing

1 INTRODUCTION

Gender biases in language are studied from many different perspectives. Sociolinguistic studies report how language use differs between men and women (e.g., women tend to have a richer vocabulary, use typical grammatical structures, and express themselves more moderately) (Lakoff, 1973; Tannen, 1990; Argamon et al., 2003). Observations that language use varies between the genders inspired author profiling studies on texts in different languages and of different genres (Koolen and van Cranenburgh, 2017; Pardo et al., 2015; Martinc et al., 2017), also in Slovene (Verhoeven et al., 2017; Škrjanec et al., 2018).[1]

The gender dimension is present as a linguistic variation in corpora and in the form of multi-layered bias, both in individual texts and in larger corpora. Research suggests that:

- The bias is manifested as lack of mentions of women: corpora often used in research contain significantly fewer female pronouns (Zhao et al., 2018) or other references to women (Caldas-Coulhard and Moon, 2010; Baker, 2010).
- Women are less often authors or editors (Hill and Shaw, 2013): only 16% of Wikipedia editors are female.
- Corpora capture stereotypical collocations (Pearce, 2008), which refer to women primarily through their reproductive function (Gorjanc, 2007) and do not associate them with (social) power (Baker, 2010).

Recent rapid developments in natural language processing (NLP) are primarily associated with the use of deep neural networks. Their use requires a representation of text in the form of numeric vectors, called word embeddings. The relations between words are expressed in the geometry of the embedded vector space: semantically related embeddings lie close in the vector space and are arranged in similar directions. This enables the study of relations beyond superficial similarities between words, e.g. through analogies such as the

1 Note that in these studies non-binary identities are not considered. Male or female gender is assigned based on, for example, author's username on social media platforms or based on other grammatical markers.

Slovenščina 2.0, 2021 (1)

relationship *Madrid:Spain* being analogous to the relationship *Paris:France* (Mikolov et al., 2013b).

As it turns out, word embeddings often contain bias, be it gender, race, or other types. Biases in word embeddings manifest through semantic associations and consequent proximities in the vector space (Mikolov et al., 2013b). Biases can be numerically evaluated by, for example, calculating cosine similarity between embeddings that describe a specific concept (e.g. gender) and potentially biased concepts. For example, Caliskan et al. (2017) show that word embeddings associate women with arts and men with science. Utilizing the aforementioned cosine similarity, a powerful approach to demonstrate potential bias in word embeddings is through a calculation of occupational analogies (Bolukbasi et al., 2016). Denoting a vector of word *w* with *v(w)*, this approach checks the existence of the following relationships between male and female word vectors: *v(man) - v(male occupation) ≈ v(woman) - v(female occupation)*. An example for Slovene is *v(moški) - v(učitelj) ≈ v(ženska) - v(učiteljica)*, where *učitelj* and *učiteljica* correspond to the masculine and feminine form of the noun for the concept (occupation) *teacher*, while *moški* and *ženska* denote *man* and *woman (the gender concept),* respectively. In case of no gender bias, the relationship between vectors for man and the masculine form of occupation and between the vector for woman and the feminine form of the same occupation would be approximately the same, as illustrated in Figure 1. However, being derived from naturally occurring text, it is not unexpected that human biases and social positions are captured in embeddings.

The illustration shows a simplified depiction of a few examples with 2-dimensional vectors. The arrows represent the difference between vectors *v(f)* and *v(m)*. The end points of arrows originating in masculine nouns for occupations represent the expected positions of equivalent feminine nouns if there were no bias.

In addition to studies that have shown the bias in word embeddings, different biases can be transferred onto algorithms for different NLP tasks, from machine translation (Prates et al., 2020; Vanmassenhove et al., 2018) to sentiment analysis (Kiritchenko and Mohammad, 2018). On the other hand, some authors (Nissim et al., 2019) warn that the analogy task's design may excessively emphasise biases.

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

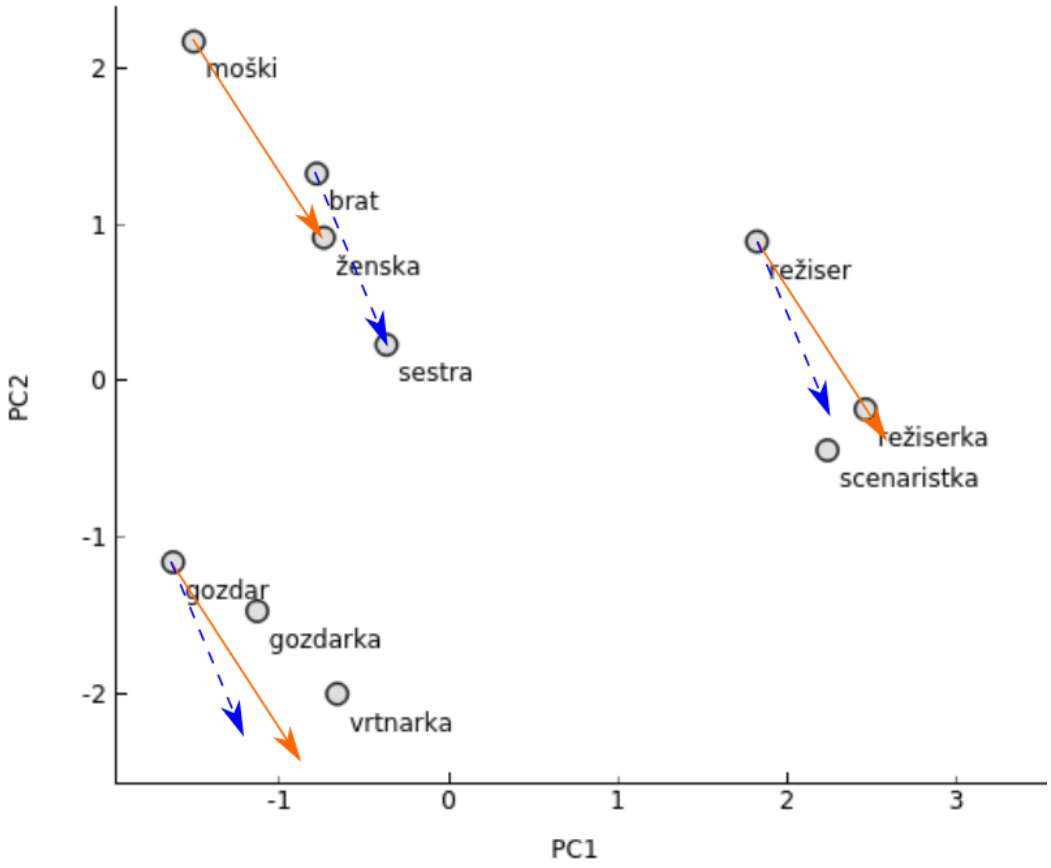


Figure 1: A simplified depiction of word vectors. The orange full arrow represents the difference between vectors for ženska [woman] and moški [man]. The blue dashed arrow represents the difference between vectors for sestra [sister] and brat [brother]. These two arrows indicate the expected (non-biased) gender difference vectors. For two male occupations, režiser [film director$_M$] and gozdar [forester$_M$], we add the gender difference vectors, and depict the resulting nearest female occupations (analogies), i.e. (gozdarka [forester$_F$] and vrtnarka [gardener$_F$]; režiserka [film director$_F$] and scenaristka [scriptwriter$_F$]). The difference to the expected non-biased point is larger for the gozdar - gozdarka pair.

Our study makes certain simplifications. First, we are not paying attention to non-binary expressions of gender, for example we do not specifically address the references such as *on/ona* or a newly proposed form introduced to be more inclusive of nonbinary gender identities *on_a* (Kern and Dobrovoljc, 2017) or noun writings of type *učitelj/učiteljica* (and *učitelj_ica*). Next, for many professions, the male form can be used as a general reference for a profession regardless of gender and we do not make any distinction between mentions of occupations when relating to a male representative or using a general mention (note also that unmarkedness of the masculine form in terms of gender is not anymore universally accepted (Kern and Dobrovoljc, 2017; Popič and

Slovenščina 2.0, 2021 (1)

Gorjanc, 2018)). As we analyse and compare the gender bias between different embedding models, these are not severe limitations, as all the embedding models are treated equally. Moreover, similar studies on languages where the gender of a noun is not expressed morphologically can run into more serious problems (see the warnings by Nissim et al. (2019)).

The main contribution of the paper is the evaluation of Slovene and Croatian word embedding models in terms of gender, which has not yet been sufficiently researched (the exception being the analysis of the Slovene w2v model in Supej et al. (2019) and Croatian evaluation of embeddings in Svoboda and Beliga (2018)). The paper extends our work (Supej et al., 2020), where we focused on quantitative evaluation and comparison of a wide range of Slovene models and different approaches to evaluation, while in this paper, we extend the work and also compare Croatian word embeddings models. The focus of the paper is to draw the attention of the developers of linguistic and technological tools (which are based on word embeddings) to the implications the usage of biased embeddings might have. Despite indirectly problematising language bias and pointing out several stereotypical associations, a detailed critical interpretation falls out of this paper's scope.

The paper is divided into further six sections. We first present related work (Section 2). Section 3 describes Slovene and Croatian lists of male and female occupations and specifies the word embedding models used. In Sections 4 and 5, methodology and results are addressed, followed by a discussion in Section 6, and conclusions with plans for further work in Section 7.

2 RELATED WORK

Language corpora and datasets reflect linguistic variations (including different types of bias) in relation to social factors. NLP tools are trained on these data and can inherit the contained variations and biases. The bias in corpora can negatively impact NLP tools (Sun et al., 2019) and can perpetuate biases held towards certain groups. Word embeddings are trained on large corpora to capture syntactic and semantic relations between words and capture the expressed biases.

For instance, it has been shown that standard training data sets for part-of-speech perform better on older people's language (Hovy and Søgaard, 2015).

Garimella et al. (2019) show that a part-of-speech tagger and a dependency parser perform successfully on texts written by women, regardless of what data they had been trained on initially. On the other hand, male authors' texts are better tagged/parsed when the training data contained enough texts written by men. The success of tools such as parsers on male authors' texts may be due to the imbalances in the training data favouring male authorship. It has also been shown that NLP tools are more effective when demographic variations are considered (Volkova et al., 2013; Hovy, 2015). Hovy (2015) shows that including the information on the age and gender of authors improves the performance of three tasks in five different languages.

Biases can have negative consequences in the coreference resolution task (Zhao et al., 2018) and can perpetuate biases held towards certain groups (see examples in Zhao et al., 2017). In the context of texts on mental illness, Hutchinson et al. (2020) note that topics such as gun violence, homelessness, and addiction are over-represented, leading to disability topics receiving particularly negative scores in sentiment analysis tasks. Besides the aspects above, some authors call the attention to the effect biases can have on detection tools. For example, misogyny detection models may attribute high scores to non-misogynous texts simply because the latter contain the so-called identity terms, i.e. terms associated with misogyny (Nozza et al., 2019). In sum, the interplay of bias and NLP is an important and interesting field receiving increasing attention, notably regarding word embeddings, as explained next.

In terms of word embeddings, researchers have studied bias by investigating the proximity of gender-related words to other words in the vector space. For example, Garg et al. (2018) show that the adjective *honourable* lies closer to the word *man* than to the word *woman*. Second, biases are reflected in analogies, e.g. Bolukbasi et al. (2016) show that the embedding space solution of the analogy *man:computer programmer* ≈ *woman:x* is *x* = *homemaker*. Nissim et al. (2019) warn that such analogies overemphasise the practical impact of the biases.

As already mentioned, gender bias in word embeddings is often studied on analogies of occupations, which is also our study's case. In morphologically rich languages, such as Slovene and Croatian, the gender of words is expressed morphologically. Therefore, the result of the gender analogy is expected to be

Slovenščina 2.0, 2021 (1)

the female form of the male variant of the occupation (and vice versa). Svoboda and Beliga (2018) included masculine and feminine versions of job positions in Croatian as one of the evaluation aspects of Croatian word2vec and fastText word embeddings. Preliminary research on word2vec embeddings in Slovene (Supej et al., 2019) showed that the analogy task's accuracy is reasonably high both when attempting to find the female and the male equivalent of an occupation. Results nevertheless reflect gender biases: the first result of the analogy *woman:secretary* ≈ *man:x* is x = *boss*, while the first ten results of different analogies indicate other gender inequalities: the association of women with house chores and men with occupations of a higher status etc. In the work of Supej et al. (2020) that we extend in this paper, different word2vec, fastText and ELMo embeddings are compared on Slovene pairs of male and female occupations.

As tools based on biased word embeddings may reinforce biases (Zhao et al., 2017), many research groups focused on *debiasing* word embeddings: the main goal of such algorithms is to prevent language models from reproducing racist, sexist or in other ways harmful content. Debiasing also has other advantages – it has been shown that debiasing contributes to correct coreference resolution (Zhao et al., 2018). Some examples of these methods are equalising the distances between gender-specific words and occupations (Bolukbasi et al., 2016; Bordia and Bowman, 2019), inserting additional restrictions into the training corpus (e.g. ensuring equal representation of occupational activities between the genders in the training data) (Zhao et al., 2017), removing texts that cause bias (Brunet et al., 2019), and training gender-neutral word embeddings (Zhao et al., 2018). Schick et al. (2021) recently proposed a self-diagnosis and self-debiasing model where large language models examine their outputs regarding the potential presence of undesirable attributes. They introduced a debiasing algorithm that reduces the likelihood of a model producing biased text. Moreover, researchers recently also focused on methods for debiasing sentence representations, addressing the difficulty of retraining models that are often proposed in debiasing research (retraining models like BERT and ELMo often proves infeasible in practice) (Liang et al., 2020). Gonen and Goldberg (2019) caution that many debiasing methods only conceal bias, which continues to be present in the embeddings, and that many metrics used in the debiasing

research have only positive predictive ability (i.e. they can detect the presence of bias but not its absence). On the other hand, studies such as Hirasawa and Komachi (2019) show that debiasing improves multimodal machine translation, thereby underlining the promising future of this research field. In our study, we do not aim to debias embeddings but only compare different embedding approaches in Slovene and Croatian concerning their gender bias.

3 DATA

In this section, we first present the lists of occupations in Slovene and Croatian we used to analyse gender biases, followed by the embedding models.

3.1 List of occupations

We first describe the list of occupations we collected for Slovene, followed by its equivalent in Croatian. Our selection of occupations in Slovene is based on the Standard Classification of Occupations (Vlada RS, 1997), based on the *International Standard Classification of Occupations*. Most occupations in this classification are multi-word expressions (e.g. *upravljalec/upravljalka metalurškega žerjava* [en. *metallurgical crane operator*]), which are less suitable for computation with embeddings due to their specificity and length. To calculate analogies, we limit our approach to single-word occupations. The complete list of single-word occupations in Slovene includes 422 male/female occupation pairs, further reduced in line with the following criteria:

1. An occupation has to exist both in female and male grammatical gender (gender-neutral words such as *pismonoša* [en. *postman*] are not included in the list).
2. An occupation as a common noun occurs at least 500 times in the Corpus of Written Standard Slovene *Gigafida 2.0* (2020).
3. When a more established version of the occupation exists, we manually add a synonym with the same root (e.g. in the case of *fotografka,* an arguably more established *fotografinja* was added [en. *photographer*]). When calculating analogies, the form more frequent in the corpora is inserted at the input, but all synonyms (if they appear among the results) are considered a correctly solved analogy.

Slovenščina 2.0, 2021 (1)

4. If the standard classification does not include the female (e.g. *dramatik* [en. *playwright*]) or male variant (e.g. *prostitutka* [en. *prostitute*]) of the occupation, the missing version is manually added if it exists and appears in the Gigafida corpus (e.g. there are no established words for female and male versions of *postrešček* [en. *porter*] and *hostesa* [en. *hostess*], respectively).

5. Occupations where either the female or the male occupation variant is a homograph (e.g. *detektivka* [en. *detective*] also denotes a detective novel) or where an occupation could be associated with a context unrelated to occupations (e.g. *čarovnik/čarovnica* [en. *wizard/witch*]), were excluded from the final set of occupations. Likewise, we filtered out occupations that are also proper names, such as *kovač* [en. *blacksmith*]; for differentiating between common nouns and proper names Sloleks 2.0 (Dobrovoljc et al., 2019) was used. The final list contains 234 occupation pairs and is freely accessible in the CLARIN repository[2].

For Croatian, we compiled a list of occupations from two existing sources. The first source contains occupations from the word analogy dataset by Svoboda and Beliga (2018). It consists of 109 pairs of single-word occupations. The second source is ESCO (European Skills, Competences, Qualifications and Occupations)[3] and lists 2942 occupations in male and female form. Similar to the Slovene list of occupations, most of the classifications from ESCO are multi-word expressions, e.g. *špediterski službenik / špediterska službenica za uvoz i izvoz riba, rakova i mekušaca* [en. *import-export specialist in fish, crustaceans and molluscs*]. After removing all multi-word occupations, the ESCO source contains 309 pairs of single-word occupations. The final, combined list from both sources, filtered to remove duplicates, contains 375 occupation pairs.

3.2 Word embedding models

Different configurations of word embeddings for Slovenian and Croatian were used in the experimental phase. We first list the Slovene embedding models followed by the Croatian ones.

2 http://hdl.handle.net/11356/1347

3 https://ec.europa.eu/esco/portal

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

3.2.1 Slovene word embedding models

We analyse two non-contextual embedding models, fastText and word2vec, and the ELMo contextual model.

- fastText (Bojanowski et al., 2017):
  - 100-dimensional vectors, trained on Gigafida 2.0 in the EU EMBEDDIA[4] project,
  - 300-dimensional vectors, trained as above,
  - 100-dimensional word vectors from the Sketch Engine portal (*word*),
  - 100-dimensional word vectors from the Sketch Engine portal, where vectors are embeddings of word lemmas,
  - 100-dimensional CLARIN.SI-embed.sl vectors (Ljubešić and Erjavec, 2018), and
  - 300-dimensional vectors from the fastText.cc portal;
- word2vec (Mikolov et al., 2013a): 256-dimensional vectors, trained for the needs of the Kontekst.io portal (Plahuta, 2020); available at request[5];
- ELMo (Peters et al., 2018): 1024-dimensional vectors, contextual embeddings built in the EU EMBEDDIA project, trained on Gigafida (Ulčar, 2019). Contextual embeddings produce a different vector for each occurrence of the word based on its context. We computed word vectors from sentences in Slovene Wikipedia. To get a single representation for each word, comparable to other embeddings, for each of the 200,000 most common words, we calculated the centroid vector of all word occurrences. Several different types of vectors were used:
  - vectors from the output of the first (CNN) layer of the network that is context-independent (i.e. *layer 0*),

4 http://embeddia.eu/

5 https://kontekst.io/kontakt

Slovenščina 2.0, 2021 (1)

- vectors from the output of the second (first LSTM) layer of the network that is context-dependent (i.e. *layer 1*),
- vectors from the output of the third (second LSTM) layer of the network that is context-dependent (i.e. *layer 2*).

3.2.2 Croatian word embedding model

For the Croatian language, we analyse several non-contextual embedding models:

- fastText (Bojanowski et al., 2017):
  - 100-dimensional vectors, trained in the EU EMBEDDIA project,
  - 300-dimensional vectors, trained as above,
  - 100-dimensional CLARIN.SI-embed.hr vectors of words and lemmas (Ljubešić, 2018),
  - 300-dimensional vectors from the fastText.cc portal.

4 EVALUATION METHODOLOGY

To assess the gender bias for each of the embedding models and each occupation, we calculated occupational analogies in four ways. However, the core analogy computation is the same in all cases: for every occupation of a masculine grammatical gender O_m, we search for a feminine noun equivalent O_f. The following vector is calculated:

$$v(d) = v(O_m) - v(m) + v(f),$$

where $v(m)$ is the male vector, and $v(f)$ is the female vector. If there were no gender biases, $v(d)$ would be equal or very similar to $v(O_f)$. For every vector $v(d)$, we find N closest word vectors according to the cosine similarity (we use N = 1, 5, or 10). When searching for closest words, all words appearing in the embeddings are considered, except for the words *man*, *woman*, the word O_m, and the words containing non-alphabetic characters (numbers, hyphens, punctuation etc.). If the word O_f is located among the N-closest words, we consider the analogy correct; else it is marked as incorrect. We convert all letters to lowercase: e.g. the words *Zdravnik*, *zdravnik* and *ZDRAVNIK* are

all converted to *zdravnik* and thus considered the same word. The process is repeated for each female variant of an occupation O_f where we look for the male equivalent O_m. Here, the vector *v(d)* is calculated as:

$$v(d) = v(O_f) - v(f) + v(m).$$

When looking for closest words, O_f is omitted from the set of words, just as O_m was ignored before. The final result represents the proportion of correctly determined cases. The metric is called *precision at N* (*P@N*). A higher *N* allows for finding additional closest hits in the vector space.

Two approaches were used to determine the baseline male vector *v(m)* and female vector *v(f)*:

- The first approach defines *m* simply as the word *man* and *f* as *woman* (in Slovene corresponding to *moški* and *ženska* and in Croatian to *muškarac* and *žena)*.
- In the second approach, similarly to Bolukbasi et al. (2016), the difference *v(f) –v(m)* or *v(m) –v(f)* is defined as the average difference of vectors of word pairs which refer specifically to a woman or man (Table 1).

Table 1: *Inherently male-female word pairs in Slovene (left) and Croatian (right)*

Slovene male-female word pairs		**Croatian male-female word pairs**	
m	***f***	***m***	***f***
moški [man]	ženska [woman]	muškarac [man]	žena [woman]
gospod [sir]	gospa [madam]	gospodin [sir]	gosopođa [madam]
fant [boy]	dekle [girl]	momak [boy]	djevojka [girl]
deček [boy]	deklica [girl]	dječak [boy]	djevojčica [girl]
brat [brother]	sestra [sister]	brat [brother]	sestra [sister]
oče [father]	mati [mother]	otac [father]	majka [mother]
sin [son]	hči [daughter]	sin [son]	kći [daughter]
dedek [grandfather]	babica [grandmother]	djed [grandfather]	baka [grandmother]
mož [husband]	žena [wife]	suprug [husband]	supruga [wife]
on [he]	ona [she]	on [he]	ona [she]
fant [boy]	punca [girl]	tata [dad]	mama [mum]
stric [uncle]	teta [aunt]		

Slovenščina 2.0, 2021 (1)

When searching for the N closest words, we also tested lemmatisation's influence: in this case, all words in word embeddings were lemmatised using the LemmaGen[6] tool. By doing so, the effect of different word forms stemming from, e.g. conjugation and declination, was offset: for example, word forms *zdravnico* and *zdravnice* are considered a single near word since they share the same lemma *zdravnica* [doctor$_F$].

5 RESULTS

We present the results showing biases in all embeddings described in Section 3. We use the *P@N* measure, where *N* equals 1, 5, or 10. Some of the occupations from our list are not covered by all word embeddings, i.e. there is no word vector for them. Any example where the searched-for word is not among the top *N* closest words is counted as incorrect, even if the searched-for word does not appear in the embeddings. In cases where the embeddings do not cover the input occupation, and we cannot calculate the vector *v(d),* we dismiss all such examples so that they do not affect the final result. The reader, interested in the results where non-covered examples are also considered, is referred to our conference paper (Supej et al., 2020).

The results for Slovene analogies are presented in Table 2 and for the Croatian analogies in Table 3. Results for experiments where we have a masculine expression for the occupation O_m as the input, and we search for the equivalent feminine expression of the same occupation O_f, are shown in the rightmost columns (*m* input) for each language. Results, where we have O_f as the input and search for O_m, are shown in leftmost columns (*f* input) for each language. As explained in Section 4, we tested different approaches. The approaches where we lemmatised all the words or used the average difference of vectors of pairs of words from Table 1 generally perform better (i.e. they express lower gender bias). These two options have the suffixes *lem* and *avg* appended in the tables, respectively. In this section, we only show the results for applying both of these options (we do not apply lemmatisation to fastText (lemma) embeddings as they are already lemmatised). Full results are presented in Appendix A in Table 8 for Slovenian and in Table 9 for Croatian.

6 https://github.com/vpodpecan/lemmagen3/

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

Table 2: *Results for all Slovenian embeddings*

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
ELMo Embeddia	1024D l0 lem avg	0.907	0.933	0.947	0.370	0.398	0.403
	1024D l1 lem avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l2 lem avg	0.880	0.933	0.933	0.376	0.398	0.398
fastText.cc	300D lem avg	0.613	0.884	0.948	0.655	0.755	0.764
fastText Embeddia	100D lem avg	0.906	0.971	0.976	0.677	0.720	0.724
	300D lem avg	**0.947**	**0.976**	**0.982**	0.685	0.720	0.724
fastText CLARIN.SI-embed.sl	100D lem avg	0.839	0.940	0.950	**0.761**	**0.880**	**0.902**
fastText Sketch Engine (word)	100D lem avg	0.930	0.962	0.973	0.725	0.781	0.785
fastText Sketch Engine (lemma)	100D avg	0.673	0.931	0.960	0.598	0.786	0.821
word2vec Kontekst.io	256D lem avg	0.679	0.853	0.872	0.407	0.550	0.593

Note. Results for each approach, where we have a feminine word for occupation on the input (*f* input), and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input), and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 3: *Results for all Croatian embeddings*

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText.cc	300D lem avg	0.731	0.939	0.954	0.546	0.637	0.644
fastText Embeddia	100D lem avg	0.905	0.941	0.968	0.625	0.666	0.672
	300D lem avg	**0.923**	**0.982**	**0.986**	0.631	0.675	0.678
fastText CLARIN.SI-embed.hr (word)	100D lem avg	0.907	0.930	0.944	**0.673**	**0.746**	**0.754**
fastText CLARIN.SI-embed.hr (lemma)	100D avg	0.244	0.678	0.826	0.266	0.521	0.588

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

The results show that both lemmatisation of the words and using the average of several inherently male or female words for male and female vectors improve the reported scores. Applying both approaches gives the best results in most cases. For finding the closest *N* words, we have also tried the CSLS

Slovenščina 2.0, 2021 (1)

measure (Cross-Domain Similarity Local Scaling) (Conneau et al., 2018) instead of the cosine similarity. This measure avoids the problem of hubness in the search for nearest neighbours. Namely, some words (called hubs in the nearest neighbour graph representation) may be nearest neighbours of many other words, while others are nearest neighbours of no other word (outliers). CSLS computes nearest neighbours in both directions and largely avoids the problem of hubness. For the experiments with O_f on the input and searching for O_m, there is no significant difference in results between the cosine similarity and CSLS. For the experiments with O_m on the input and searching for O_f, using CSLS gives lower precision than the cosine similarity. This is especially the case where we used the words "man" and "woman" for vectors *v(m)* and *v(f)*. When using averages of several inherently male and female words for vectors *v(m)* and *v(f)*, the difference in precision between the cosine similarity and CSLS is smaller, but the cosine similarity still outperforms CSLS.

We give a more detailed discussion of the results for each approach in the next section. We only present the results of the cosine similarity measure.

6 DISCUSSION

In the case of Slovene word embeddings, the fastText CLARIN.SI-embed.sl embeddings reach the highest precision in the analogy task for male versions of occupations at the input (Table 2). When there are female versions of occupations at the input, the embedding model reaching the highest precision is fastText Embeddia. Similar results are observed for Croatian embeddings (Table 3). Lemmatisation of the output and averaging several inherently male and female words for vectors *v(m)* and *v(f)* (instead of using only the embeddings for *woman* or *man*) improves the precision in the analogy task for different models and different input data. As described in Section 5, we dismiss the examples where the embeddings do not cover the input occupation. If we do not dismiss these examples but instead count them as incorrect, the share of occupations covered by the embeddings has the largest effect on the score. The results for Slovene can be found in our paper (Supej et al., 2020). The fastText CLARIN.SI embeddings would then score the best, as these embeddings cover the occupations best. This is especially important for the female occupations since they have much lower coverage than male occupations.

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

Results in Table 2 and Table 3 have been filtered, so that the words *man, woman* and the occupation on the input are removed from the list of analogy results, as explained in Section 4. With unfiltered results, the input occupation is often the result of the analogy task (Table 4). For more detailed results (not only with lemmatisation and using several inherently male and female words for *v(m)* and *v(f)*) see Table 10 in Appendix A.

With the fastText Embeddia model, we reach similar results using 100- and 300-dimensional vectors (see Table 2 and Table 3). Other embeddings are not directly comparable with regards to dimensionality as they were trained on different resources. However, corpora used to train the embeddings play a more important role than the number of dimensions. The FastText Embeddia model in Table 4 shows that dimensionality plays a role in determining how often the input occupation is the result of the analogy. In a different setup, when considering the occupations that are not covered in the embeddings, dimensionality strongly influences the results (Supej et al., 2020).

Table 4: *Share of cases where the result of the analogy with the highest cosine similarity is the input occupation itself - before filtering is done to produce the results in Table 2 and Table 3 (both male to female and female to male analogies)*

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
ELMo Embeddia	1024D l0 lem avg	0.547			
	1024D l1 lem avg	0.423			
	1024D l2 lem avg	**0.064**			
fT fastText.cc	300D lem avg	0.831	fT fastText.cc	300D lem avg	0.672
fT Embeddia	100D lem avg	0.143	fT Embeddia	100D lem avg	**0.094**
	300D lem avg	0.419		300D lem avg	0.352
fT CLARIN.SI-embed.sl (word)	100D lem avg	0.316	fT CLARIN.SI-embed. hr (word)	100D lem avg	0.103
fT Sketch Engine (word)	100D lem avg	0.096			
fT Sketch Engine (lemma)	100D avg	0.803	fT CLARIN.SI-embed.hr (lemma)	100D avg	0.837
w2v Kontekst.io	256D lem avg	0.483			

Note. The number of all cases is 468 (from 234 occupation pairs) for Slovene and 750 (from 375 occupation pairs) for Croatian.

Slovenščina 2.0, 2021 (1)

The coverage of masculine occupations is higher than that of feminine occupations in all word embedding models (Table 5). FastText CLARIN.SI-embed.sl word embeddings achieve the highest coverage of female occupations, while ELMo word embeddings contained only 75 of the 234 female occupations. As explained in Section 3.2.1, ELMo embeddings are limited to only 200,000 most common words in Wikipedia; therefore, we have significantly lower coverage of occupations for ELMo. For comparison, other word embedding models cover around 1 million words. Masculine occupations that do not appear in the embeddings are typically occupations associated with women (e.g. male variants of *seamstress* and *cosmetician*, in Slovene *šiviljec* and *kozmetik*, respectively). Likewise, feminine occupations not present in the embeddings are traditionally male occupations (e.g. embedding models do not contain female variants of occupations like *auto mechanic* and *carpenter* (in Slovene *avtomehaničarka* and *tesarka*, respectively), or occupations that have been culturally taken up exclusively by men, e.g., *nadškof* (en. *archbishop*). Poor representation of female occupations can also be attributed to other factors — Zhao et al. (2018) report that the mentions referring to men are more likely to contain a job title compared to female mentions.

Table 5: *Coverage of male (m) and female (f) occupations from the list in different embeddings as a ratio between covered occupations and all occupations*

Slovene embeddings	*m*	*f*	Croatian embeddings	*m*	*f*
ELMo	0.774	0.321			
fastText cc	0.979	0.739	fastText cc	0.848	0.527
fastText Embeddia	0.991	0.726	fastText Embeddia	0.856	0.594
fastText CLARIN.SI-embedd.sl	**1.000**	**0.932**	fastText CLARIN.SI-embedd.hr (word)	0.914	**0.722**
fastText Sketch Engine (word)	0.996	0.791	fastText CLARIN.si-embedd.hr (lemma)	**0.955**	**0.722**
fastText Sketch Engine (lemma)	**1.000**	0.863			
word2vec Kontekst.io	0.987	0.667			

Nissim et al. (2019) claim that most studies exaggerate biases pointed out by analogy tasks. The design of these studies excludes the input occupation from the possible results, even if the calculations could lead to this exact occupation to have the highest cosine similarity and hence appear in the results. This criticism is more relevant for English studies as in Slovene the gender in

occupations is for the most part expressed by word morphology. Even though we omitted the input occupations from the results, which is a standard practice when calculating analogies, we analysed the results before this filtering. Analysis of the results showed that the input occupation is indeed often the result with the highest cosine similarity (Table 4), varying significantly between different models.

When manually comparing the results of different models from Tables 2 and 3, we also notice several differences between the models. In the case of ELMo and word2vec models, the outputs are largely occupations. The results of the analogy task in the case of fastText Embeddia, CLARIN.SI-embed.sl and Sketch Engine (word) are occupations, as well as words related to the occupation on the input, or words that share the same root as the input occupation. Results of the fastText.cc and Sketch Engine (lemma) models are typically words sharing the root with the input occupation.

Analogy results are interesting from a semantic point of view. The first results of the analogy task (Slovene "fastText Embeddia 100D lem avg") *ženska:krojačica :: moški:x* being *x=krojač* [en. *woman:tailor*$_F$ *:: man:tailor*$_M$] and *ženska:šivilja :: moški:x* being *x=krojač* [en. *woman:seamstress :: man:tailor*] are interesting. For example, while word embedding of *šiviljec* [en. *seamster*] is not available, *krojač* [en. *tailor*], a semantically linked one, from another morphological word family is. Another interesting element is illustrated by one of the results of the analogy: *ženska:manekenka :: moški:x* where *x=nogometaš* [en. *woman:model :: man:footballer*] (Croatian "fastText Embeddia 100D lem avg"). While *model* and *footballer* are not corresponding to the same professions, this result is an indication that female models and male footballers appear in similar textual contexts. It would be interesting to investigate those contexts further (e.g. both occupations represent desirable identities, such as being beautiful, rich, famous, successful).

There are indeed more examples where results of certain analogies (especially in the case of "word2vec Kontekst.io lem avg model") are not linked to the input occupation or are stereotypical. For example, the results of the analogy *moški:rudar :: ženska:x* in the aforementioned w2v model are, e.g. *barbika* [en. *barbie*], *klovnesa* [en. *clown*$_F$], *čarovnica* [en. *witch*], *lutka* [en. *doll*], *prostitutka* [en. *prostitute*$_F$], *akrobatka* [en. *acrobat*$_F$], *najstnica* [en.

Slovenščina 2.0, 2021 (1)

teenager$_F$], *opica* [en. *monkey*], *princeska* [en. *princess*], *striptizeta* [en. *stripper*$_F$]. The case of stereotypical analogies in the w2v model is pointed out by Supej et al. (2019).

As part of the analysis, a frequency list of analogy results for female and male input occupations was compiled for each word embedding model (only the *lem avg* configuration of the models was taken into account) (see Table 6 for Slovene and Table 7 for Croatian).

The most frequently occurring words mostly follow the pattern that for a male occupation on the input, a female occupation is expected on the output. Presented Slovene embedding models follow this pattern; in the case of the Croatian embeddings, there are several examples among the frequently occurring words that do not follow the pattern: in the "fastText cc lem avg" with a female occupation on the input, there are several frequently occurring female occupation variants also on the output, e.g. *ethicist*, *biologist* (*etičarka*, *biologinja*, respectively). For *etičarka,* it is possible that this result is influenced by other similar words (e.g. *kozmetičarka*), as fastText models consider subword information. The most frequently occurring words are primarily occupations but not always – for example, female Scottish national (*Škotkinja*) and *father* (*otac*) frequently appear in the Croatian "fastText cc lem avg" model while one of the frequent words in the Slovene "word2vec Kontekst.io lem avg" is *korenjak* (denoting a brave man).

In Slovene word embeddings, we notice a pattern of the most frequently occurring feminine occupations/words appearing more often than the most frequently occurring male occupations in the "ELMo l2 lem avg" and "w2v Kontekst.io lem avg" models. Similar is observed for Croatian models presented in Table 7; however, the most frequently occurring words appear less often than in the Slovene embeddings. One possible explanation is that the models mentioned above contain fewer word embeddings than some other models (200,000 or approximately 600,000 for each model). Both models exhibit a lower representation of the female versions of occupations in the embeddings. Occupations that nevertheless appear in the embeddings, therefore, reappear more often. There are overall more male occupations in the embeddings, possibly causing individual male occupations to come up less frequently than female ones.

Table 6: *Most common words that appear among the top 10 results of the analogy task (that is, among the 10 closest words to the searched-for term, based on the cosine similarity measure) for selected Slovene embedding models*

ELMo Embeddia l2 lem avg				**fastText CLARIN.SI lem avg**				**word2vec Kontekst.io lem avg**			
***m* input**		***f* input**		***m* input**		***f* input**		***m* input**		***f* input**	
Result	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***
bolničarka [nurse]	47	geograf [geographer$_M$]	9	šivilja [seamstress]	15	mizar [carpenter$_M$]	11	kuharica [cook$_F$]	44	ortoped [orthopedist$_M$]	14
biokemičarka [biochemist$_F$]	39	politolog [political scientist$_M$]	8	ključavničarka [locksmith$_F$]	11	biology [biologist$_M$]	10	gospodinja [homemaker$_F$]	38	pisatelj [writer$_M$]	14
frizerka [hairdresser$_F$]	39	biolog [biologist$_M$]	7	inštalaterka [installer$_F$]	9	ključavničar [locksmith$_M$]	9	šivilja [seamstress]	33	kardiolog [cardiologist$_M$]	13
trgovka [salesperson$_F$]	39	dramaturg [playwright$_M$]	7	keramičarka [ceramist$_F$]	9	zgodovinar [historian$_M$]	9	frizerka [hairdresser$_F$]	32	nevrolog [neurologist$_M$]	13
čistilka [cleaner$_F$]	34	književnik [writer$_M$]	7	filologinja [philologist$_F$]	8	internist [internist$_M$]	8	kozmetičarka [cosmetician$_F$]	30	urolog [urologist$_M$]	13
znanstvenica [scientist$_F$]	34	scenarist [screenwriter$_M$]	7	oftalmologinja [ophthalmologist$_F$]	8	režiser [director$_M$]	8	čistilka [cleaner$_F$]	29	psihiater [psychiatrist$_M$]	12
kuharica [cook$_F$]	33	animator [animator$_M$]	6	filozofinja [philosopher$_F$]	7	arheolog [archeologist$_M$]	7	fotografinja [photographer$_F$]	29	ekolog [ecologist$_M$]	11
geologinja [geologist$_F$]	30	esejist [essayist$_M$]	6	geofizičarka [geophysicist$_F$]	7	natakar [waiter$_M$]	7	zdravnica [doctor$_F$]	29	hišnik [janitor$_M$]	11
perica [laundress]	28	etnolog [ethnologist$_M$]	6	kmetica [farmer$_F$]	7	pisatelj [writer$_M$]	7	služkinja [maid]	26	biolog [biologist$_M$]	10
služkinja [maid]	28	fotograf [photographer$_M$]	6	nevrokirurginja [neurosurgeon$_F$]	7	primarij [senior doctor$_M$]	7	trgovka [salesperson$_F$]	26	korenjak [brave man]	10
biologinja [biologist$_F$]	27	ilustrator [illustrator$_M$]	6	strugarka [worker using a planer machine$_F$]	7	stomatolog [stomatologist$_M$]	7	slikarka [painter$_F$]	25	maneken [model$_M$]	10
gospodinja [homemaker$_F$]	26	lutkar [puppeteer$_M$]	6	geologinja [geologist$_F$]	6	tesar [carpenter$_M$]	7	tajnica [secretary$_F$]	25	režiser [director$_M$]	10
matematičarka [mathematician$_F$]	26	paleontolog [paleontologist$_M$]	6	hematologinja [hematologist$_F$]	6	fotoreporter [photojournalist$_M$]	6	veterinarka [veterinarian$_F$]	25	akademik [academic$_M$]	9
mikrobiologinja [microbiologist$_F$]	26	pravnik [jurist$_M$]	6	kardiologinja [cardiologist$_F$]	6	gostilničar [innkeeper$_M$]	6	znanstvenica [scientist$_F$]	25	akademski slikar [academic painter$_M$]	9
arheologinja [archeologist$_F$]	25	režiser [director$_M$]	6	paleontologinja [paleontologist$_F$]	6	kardiolog [cardiologist$_M$]	6	socialna delavka [social worker$_F$]	24	glasbenik [musician$_M$]	9

Slovenščina 2.0, 2021 (1)

Table 7: *15 most common words that appear among the top 10 results of the analogy task (that is, among the 10 closest words to the searched-for term, based on the cosine similarity measure) for selected Croatian embedding models*

ELMo Embeddia l2 lem avg				**fastText cc lem avg**				**fastText CLARIN.SI-embedd.hr (word) lem avg**			
***m* input**		***f* input**		***m* input**		***f* input**		***m* input**		***f* input**	
Result	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***
krojačica [tailor$_F$]	34	povjesničar [historian$_M$]	10	kemičarka [chemist$_F$]	12	etičarka [ethicist$_F$]	8	krojačica [tailor$_F$]	31	znanstvenik [scientist$_M$]	16
automehaničarka [auto mechanic$_F$]	29	konobar [waiter$_M$]	10	vještakinja [expert$_F$]	11	otfamologinja [ophthalmologist$_F$]	7	automehaničarka [auto mechanic$_F$]	23	biology [biologist$_M$]	16
zavarivačica [welder$_F$]	20	biolog [biologist$_M$]	9	fizičarka [physicist$_F$]	10	redatelj [director$_M$]	6	zavarivačica [welder$_F$]	22	profesor [professor$_M$]	9
keramičarka [ceramist$_F$]	16	umjetnik [artist$_M$]	8	biokemičarka [biochemist$_F$]	10	glumac [actor$_M$]	6	šivačica [seamstress]	18	povjesničar [historian$_M$]	9
kemičarka [chemist$_F$]	15	sociolog [sociologist$_M$]	8	vozačica [driver$_F$]	9	biologinja [biologist$_F$]	6	keramičarka [ceramist$_F$]	18	konobar [waiter$_M$]	9
biokemičarka [biochemist$_F$]	15	fizioterapeut [physiotherapist$_M$]	8	pravnica [jurist$_F$]	9	paleografkinja [paleographer$_F$]	5	soboslikarica [painter-decorator$_F$]	17	genetičar [geneticist$_M$]	9
šivačica [seamstress]	14	redatelj [director$_M$]	7	frizerka [hairdresser$_F$]	9	ihtiologinja [ichthyologist$_F$]	5	biokemičarka [biochemist$_F$]	16	redatelj [director$_M$]	8
spremačica [maid]	14	poslovođa [manager$_{F/M}$]	7	masažerka [massage therapist$_F$]	8	suscenarist [co-screenwriter$_M$]	4	kemičarka [chemist$_F$]	15	poslovođa [manager$_{F/M}$]	8
čistačica [cleaner$_F$]	13	paleontolog [paleontologist$_M$]	7	tehničarka [technician$_F$]	7	scenografkinja [scenographer$_F$]	4	genetičarka [geneticist$_F$]	13	policajac [police officer$_M$]	8
genetičarka [geneticist$_F$]	13	književnik [writer$_M$]	7	političarka [politician$_F$]	7	otac [father]	4	cvjećarka [florist$_F$]	12	zaposlenik [employee$_M$]	7
fizičarka [physicist$_F$]	13	geologinja [geologist$_F$]	7	matematičarka [mathematician$_F$]	7	književnik [writer$_M$]	4	biofizičarka [biophysicist$_F$]	12	umjetnik [artist$_M$]	7
astrofizičarka [astrophysicist$_F$]	13	dramaturg [playwright$_M$]	7	lutkarica [puppeteer$_F$]	7	dopukovnik [lieutenant colonel$_M$]	4	znanstvenica [scientist$_F$]	11	sociolog [sociologist$_M$]	7
šnajderica [seamstress]	12	znanstvenik [scientist$_M$]	6	glumica [actor$_F$]	7	daktilografkinja [typist$_F$]	4	geologinja [geologist$_F$]	11	snimatelj [cameraman]	7
mehaničarka [mechanic$_F$]	12	zaštitar [security guard$_M$]	6	trgovkinja [salesperson$_F$]	6	astrobiologinja [astrobiologist$_F$]	4	tehničarka [technician$_F$]	10	satnik [captain$_M$]	7
informatičarka [computer scientist$_F$]	12	sociologinja [sociologist$_F$]	6	terapeutkinja [therapist$_F$]	6	škotkinja [Scottish national$_F$]	3	mehaničarka [mechanic$_F$]	10	porter [doorkeeper$_M$]	7

In the case of the Slovene "ELMo l2 lem avg" and "w2v Kontekst.io lem avg" models, occupations of a lower social class (*čistilka* [en. *cleaner*$_F$], *perica* [en. *laundress*], *gospodinja* [en. *homemaker*$_F$]), as well as archaic occupations with women in inferior roles (*služkinja* [en. *maid*]) are observed among the frequent analogy results of female grammatical gender. Socially inferior occupations are rare among the most frequent male analogies. There are less socially inferior occupations observed among the Croatian results (exceptions being, e. g., the female variants of *cleaner* and *maid* (*čistačica* and *spremačica*, respectively) in the "ELMo Embeddia l2 lem avg" model).

We observed that certain words (especially female occupations) appear among the results despite being semantically unrelated to the input occupation. Several analogy results (especially in the case of a typical male occupation on the input) are unrelated to the input occupation (e.g. *bolničarka* [en. nurse$_F$] is the first result of the analogy *moški:rudar :: ženska:x* [en. *man:miner :: woman:x*] and *šivilja* [en. *seamstress*] the first result of the analogy *moški:avtomehanik :: ženska:x* [en. *man:auto mechanic :: woman:x*] in the Slovene model "fastText Embeddia 100D lem avg"). One explanation is that certain word embeddings are more "central" than the others and, therefore, the closest neighbour of many other words. To check if this explanation is true, instead of the cosine similarity measure, we used the CSLS measure (Conneau et al., 2018) that considers the shared distances of *N* closest neighbours. We observed that the precision is worse when using the CSLS measure than the cosine similarity (Section 5), and therefore we do not report these results. However, when observing the most common words, returned as the analogy task results (Table 6 and Table 7), the distribution of the most common words is more uniform when using the CSLS measure.

Direct comparison of models between Croatian and Slovene is not possible, as the embeddings are trained on different text corpora, and the professions used for analogy calculations are not the same. However, we can notice that in Croatian the occupational gender bias in tested embeddings is slightly higher. Interestingly, the statistical data shows that the employment gap and the pay gap between women and men are lower in Slovenia compared to Croatia (Eurostat, 2021). In future, it would be interesting to study if the female employment rate and gap, as well as the gap in salaries for the same professions between countries,

Slovenščina 2.0, 2021 (1)

is correlated with the gender bias in embeddings models trained on the corresponding national languages and the changes of this correlation through time.

7 CONCLUSIONS AND FURTHER WORK

We evaluated different Slovene and Croatian word embeddings on analogies of male and female occupations (using different configurations and approaches to calculate analogies). Our focus is on the quantitative evaluation, and the results may be informative for developers of NLP tools. The lowest gender bias was obtained using the fastText embeddings. In finding female analogies (male occupation on the input), the best performing models proved to be fastText CLARIN.SI-embed.sl and fastText CLARIN.SI-embed.hr for Slovene and Croatian, respectively, while the best performing models for finding male analogies (female occupation on the input) were the respective fastText Embeddia models. The approach where averages of several inherently male and female words were used instead of using only the embeddings for woman or man improved the results. Lemmatization likewise improves the precision. With female occupations at the input, the best results (P@10) of 0.982 and 0.986 are achieved using the "fastText Embeddia 300D lem avg" models for Slovene and Croatian, respectively (the examples where the embeddings do not cover the input occupation were dismissed). With male occupations on the input, the best results of 0.902 and 0.754 are produced by the "fastText CLARIN.SI-embed.sl 100D lem avg" and "fastText CLARIN.SI-embed.hr 100D (lem) avg" (cases where the input occupation is not present among the embeddings were likewise dismissed). Lowest results for male input reflect lower coverage of female occupation equivalents in the embeddings model. The "fastText CLARIN.SI-embed.sl" and "fastText CLARIN.si-embedd.hr (lemma)" models contain the highest ratio of searched-for female and male occupations. The qualitative analysis identifies the word2vec Kontekst.io model as the model with the highest degree of gender bias in the results (stereotypically male/female occupations appearing among the results regardless of the grammatical gender of the input occupation).

In future work, we will focus on a detailed qualitative analysis and the relationship between word embeddings, language, and social power. Moreover, we will align occupations in Slovene and Croatian. Further work will also encompass an evaluation of BERT contextual embeddings and experiments in

other languages. The impact of the gender bias will be tested in predictive models on practical tasks such as the sentiment analysis.

Acknowledgments

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103, as well as project no. J6-2581. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *TEXT*, *23*, 321–346.

Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender & Language*, *4*(1), 125–149.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS'16)* (pp. 4356–4364).

Bordia, S., & Bowman, S. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop,* (pp. 7–15).

Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. S. (2019). Understanding the Origins of Bias in Word Embeddings. *Proceedings of International Conference on Machine Learning (ICML 2019)*.

Caldas-Coulhard, C. R., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, *21*(2), 99–133.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, *356*(6334), 183–186.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jegou, H. (2018). Word translation without parallel data. *Proceedings of the International Conference on Learning Representation (ICLR)*.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, T., Arhar Holdt, Š., Čibej, J., Krsnik L., & Robnik-Šikonja, M. (2019). Morphological lexicon Sloleks 2.0. CLARIN.SI. http://hdl.handle.net/11356/1230

Eurostat (2021). Gender statistics. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_statistics#Labour_market

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, *115*(16).

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. *Proceedings of the 57th Annual Meeting of the ACL* (pp. 3493–3498).

Gigafida 2.0. Retrieved from https://viri.cjvt.si/gigafida

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of NAACL-HLT 2019* (pp. 609–614).

Gorjanc, V. (2007). Kontekstualizacija oseb ženskega in moškega spola v slovenskih tiskanih medijih. In I. Novak-Popov (Ed.), *Stereotipi v slovenskem jeziku, literaturi in kulturi: zbornik predavanj 43. seminarja slovenskega jezika, literature in culture* (pp. 173–180). Ljubljana: Center za slovenščino kot drugi/tuji jezik.

Hill, B., & Shaw, A. (2013). The Wikipedia gender gap revisited: Characterising survey response bias with propensity score estimation. *PloS One*, 8.

Hirasawa, T., & Komachi, M. (2019). Debiasing Word Embeddings Improves Multimodal Machine Translation. *Proceedings of Machine Translation Summit XVII, Vol. 1* (pp. 32–42).

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP* (pp. 483–488).

Hovy, D. (2015). Demographic factors improve classification performance. *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP* (pp. 752–762).

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5491–5501).

Kern, B., & Dobrovoljc, H. (2017). Pisanje moških in ženskih oblik in uporaba podčrtaja za izražanje »spolne nebinarnosti«. Jezikovna svetovalnica. Retrieved from https://svetovalnica.zrc-sazu.si/topic/2247/pisanje-mo%C5%A1kih-in-%C5%BEenskih-oblik-in-uporaba-pod%C4%8Drtaja-za-izra%C5%BEanje-spolne-nebinarnosti

Kiritchenko, S., & Mohammad, S., (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (pp. 43–53).

Koolen, C., & van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. *Proceedings of the First Ethics in NLP workshop* (pp. 12–22).

Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2*(1), 45–80.

Liang, P. P, Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L. (2020). Towards Debiasing Sentence Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5502–5515).

Ljubešić, N., & Erjavec, T. (2018). Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1204

Ljubešić, N. (2018). Word embeddings CLARIN.SI-embed.hr 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1205

Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. (2017). PAN 2017: Author profiling - gender and language variety prediction: notebook for PAN at CLEF 2017. *Proceedings of the Conference and Labs of the Evaluation Forum*.

Mikolov, T., Corrado, G. S., Chen, K., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations* (pp. 1–12).

Mikolov, T., Yih, W-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies* (pp. 746–751).

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended Bias in Misogyny Detection. *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 149–155).

Nissim, M., van Noord, R., & van der Goot, R. (2019). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(3), 487–497.

Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, *3*(1), 1–29.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualised word representations. *Proceedings of NAACL-HLT 2018* (pp. 2227–2237).

Plahuta, M. (2020). O slovarju. Retrieved from https://kontekst.io/oslovarju

Popič, D., & Gorjanc, V. (2018). Challenges of adopting gender-inclusive language in Slovene. *Suvremena lingvistika, 44*(86), 329–350.

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, *32*, 6363–6381.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In L. Cappellato, N. Ferro, G. J. F. Jones in E. SanJuan (Eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *arXiv preprint arXiv:2103.00453*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K-W., & Wang, W. Y. (2019). Mitigating gender bias in

natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the ACL* (pp. 1630–1640).

Supej, A., Plahuta, M., Purver, M., Mathioudakis, M., & Pollak, S. (2019). Gender, language, and society: Word embeddings as a reflection of social inequalities in linguistic corpora. *Proceedings of the Slovensko sociološko srečanje 2019 – Znanost in družbe prihodnosti* (pp. 75–83).

Supej, A., Ulčar, M., Robnik-Šikonja, M., & Pollak, S. (2020). Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev. *Proceedings of the Conference on Language Technologies & Digital Humanities 2020* (pp. 93–100).

Svoboda, L., & Beliga, S. (2018). Evaluation of Croatian Word Embeddings. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 1512–1518).

Škrjanec, I., Lavrač, N., & Pollak, S. (2018). Napovedovanje spola slovenskih blogerk in blogerjev. In D. Fišer (Ed.), *Viri, orodja in metode za analizo spletne slovenščine* (pp. 356–373). Ljubljana: Znanstvena založba FF.

Tannen, D. (1990). *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine Books.

Ulčar, M. (2019). ELMo embeddings model, Slovenian. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1257

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. *Proceedings of the EMNLP* (pp. 3003–3008).

Verhoeven, B., Škrjanec, I., & Pollak, S. (2017). Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. *Proceedings of the 6th BSNLP Workshop* (pp. 119–125).

Vlada RS (1997). 1641. uredba o uvedbi in uporabi standardne klasifikacije poklicev. *Uradni list RS*, *28*, 2217. Retrieved from https://www.uradni-list.si/glasilo-uradni-listrs/vsebina?urlid=199728&stevilka=1641

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. *Proceedings of the EMNLP* (pp. 1815–1827).

Slovenščina 2.0, 2021 (1)

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the EMNLP* (pp. 2979–2989).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the NAACL-HLT* (pp. 15–20).

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

PRIMERJAVA SLOVENSKIH IN HRVAŠKIH BESEDNIH VEKTORSKIH VLOŽITEV Z VIDIKA SPOLA NA ANALOGIJAH POKLICEV

V zadnjih letih je uporaba globokih nevronskih mrež in gostih vektorskih vložitev za predstavitve besedil privedla do vrste odličnih rezultatov na področju računalniškega razumevanja naravnega jezika. Prav tako se je pokazalo, da vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Prispevek se osredotoča na evalvacijo vektorskih vložitev besed v slovenščini in hrvaščini z vidika spola z uporabo besednih analogij. Sestavili smo seznam moških in ženskih samostalnikov za poklice v slovenščini in ovrednotili spolno pristranost modelov vložitev fastText, word2vec in ELMo z različnimi konfiguracijami in pristopi k računanju analogij. Izkazalo se je, da najmanjšo poklicno spolno pristranost vsebujejo vložitve fastText. Tudi za hrvaško evalvacijo smo uporabili sezname poklicev in primerjali različne fastText vložitve.

Ključne besede: besedne vložitve, spolna pristranost, besedne analogije, poklici, obdelava naravnega jezika

Slovenščina 2.0, 2021 (1)

APPENDIX 1

We present the results, comparing different approaches described in Section 4 and Section 5. The approach where we lemmatised all the words has the suffix *lem* appended in the tables. The approach where we used the average difference of vectors of pairs of words from Table 1 has the suffix *avg* appended in the tables. The results for Slovene word embeddings are shown in Table 8, the results for Croatian word embeddings in Table 9 and the share of cases, where the input occupation is the result of the analogy task, in Table 10.

Table 8: *Results for Slovenian embeddings*

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
ELMo Embeddia	1024D l0 avg	0.707	0.933	0.947	0.166	0.359	0.387
	1024D l0	0.427	0.920	0.947	0.210	0.376	0.398
	1024D l0 lem avg	0.907	0.933	0.947	0.370	0.398	0.403
	1024D l0 lem	0.893	0.947	0.947	0.376	0.392	0.403
	1024D l1 avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l1	0.880	0.947	0.947	0.376	0.392	0.392
	1024D l1 lem avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l1 lem	0.907	0.947	0.947	0.376	0.392	0.392
	1024D l2 avg	0.880	0.933	0.933	0.376	0.398	0.398
	1024D l2	0.853	0.920	0.933	0.370	0.398	0.398
	1024D l2 lem avg	0.880	0.933	0.933	0.376	0.398	0.398
	1024D l2 lem	0.853	0.920	0.933	0.370	0.398	0.398
fastText.cc	300D avg	0.393	0.798	0.913	0.607	0.738	0.751
	300D	0.150	0.561	0.792	0.445	0.703	0.734
	300D lem avg	0.613	0.884	0.948	0.655	0.755	0.764
	300D lem	0.457	0.861	0.919	0.498	0.725	0.751
fastText Embeddia	100D avg	0.900	0.971	0.976	0.672	0.716	0.720
	100D	0.471	0.871	0.906	0.638	0.716	0.720
	100D lem avg	0.906	0.971	0.976	0.677	0.720	0.724
	100D lem	0.735	0.924	0.941	0.638	0.716	0.720
	300D avg	0.835	0.971	0.976	0.668	0.716	0.724
	300D	0.329	0.859	0.959	0.685	0.720	0.720
	300D lem avg	**0.947**	**0.976**	**0.982**	0.685	0.720	0.724
	300D lem	0.818	0.971	0.976	0.685	0.720	0.720

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText CLARIN.SI-embed.sl	100D avg	0.784	0.913	0.940	**0.761**	0.868	0.880
	100D	0.083	0.587	0.780	0.705	0.855	0.885
	100D lem avg	0.839	0.940	0.950	**0.761**	**0.880**	**0.902**
	100D lem	0.651	0.881	0.917	0.709	0.859	0.885
fastText Sketch Engine (word)	100D avg	0.886	0.962	0.973	0.717	0.768	0.777
	100D	0.211	0.757	0.908	0.691	0.768	0.777
	100D lem avg	0.930	0.962	0.973	0.725	0.781	0.785
	100D lem	0.811	0.951	0.962	0.691	0.768	0.781
fastText Sketch Engine (lemma)	100D avg	0.673	0.931	0.960	0.598	0.786	0.821
	100D	0.510	0.812	0.891	0.380	0.658	0.756
word2vec Kontekst.io	256D avg	0.679	0.853	0.872	0.407	0.550	0.593
	256D	0.365	0.590	0.718	0.251	0.489	0.515
	256D lem avg	0.679	0.853	0.872	0.407	0.550	0.593
	256D lem	0.513	0.686	0.795	0.251	0.489	0.519

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 9: *Results for Croatian embeddings*

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText.cc	300D avg	0.604	0.883	0.944	0.536	0.603	0.609
	300D	0.452	0.838	0.914	0.429	0.599	0.606
	300D lem avg	0.731	0.939	0.954	0.546	0.637	0.644
	300D lem	0.660	0.924	0.954	0.508	0.618	0.634
fastText Embeddia	100D avg	0.896	0.941	0.959	0.625	0.669	0.672
	100D	0.797	0.928	0.937	0.459	0.634	0.656
	100D lem avg	0.905	0.941	0.968	0.625	0.666	0.672
	100D lem	0.833	0.932	0.941	0.503	0.641	0.662
	300D avg	0.829	0.937	0.973	0.616	0.675	0.675
	300D	0.703	0.914	0.950	0.431	0.662	0.672
	300D lem avg	**0.923**	**0.982**	**0.986**	0.631	0.675	0.678
	300D lem	0.865	0.950	0.964	0.578	0.672	0.675

Slovenščina 2.0, 2021 (1)

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText CLARIN.SI-embed.hr (word)	100D avg	0.896	0.933	0.941	0.670	**0.749**	**0.754**
	100D	0.778	0.904	0.919	0.491	0.699	0.740
	100D lem avg	0.907	0.930	0.944	**0.673**	0.746	**0.754**
	100D lem	0.815	0.904	0.915	0.550	0.711	0.746
fastText CLARIN.SI-embed.hr (lemma)	100D avg	0.244	0.678	0.826	0.266	0.521	0.588
	100D	0.278	0.593	0.693	0.126	0.336	0.406

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 10: *Share of cases where the result of the analogy with the highest cosine similarity is the input occupation itself - before filtering is done to produce the results of Tables 2 and 3 (both male to female and female to male analogies)*

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
ELMo Embeddia	1024D l0 avg	0.547			
	1024D l0	0.547			
	1024D l0 lem avg	0.547			
	1024D l0 lem	0.547			
	1024D l1 avg	0.423			
	1024D l1	0.483			
	1024D l1 lem avg	0.423			
	1024D l1 lem	0.483			
	1024D l2 avg	**0.064**			
	1024D l2	0.088			
	1024D l2 lem avg	**0.064**			
	1024D l2 lem	0.088			
fT fastText.cc	300D avg	0.831	fT fastText.cc	300D avg	0.672
	300D	0.825		300D	0.664
	300D lem avg	0.831		300D lem avg	0.672
	300D lem	0.825		300D lem	0.664

M. ULČAR et al.: Slovene and Croatian word embeddings in terms of gender...

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
fT Embeddia	100D avg	0.143	ft Embeddia	100D avg	**0.094**
	100D	0.141		100D	**0.094**
	100D lem avg	0.143		100D lem avg	**0.094**
	100D lem	0.141		100D lem	**0.094**
	300D avg	0.419		300D avg	0.352
	300D	0.513		300D	0.441
	300D lem avg	0.419		300D lem avg	0.352
	300D lem	0.513		300D lem	0.441
fT CLARIN.SI-embed.sl (word)	100D avg	0.316	fT CLARIN.SI-embed.hr (word)	100D avg	0.103
	100D	0.310		100D	0.114
	100D lem avg	0.316		100D lem avg	0.103
	100D lem	0.310		100D lem	0.114
fT Sketch Engine (word)	100D avg	0.096			
	100D	0.135			
	100D lem avg	0.096			
	100D lem	0.135			
fT Sketch Engine (lemma)	100D avg	0.803	fT CLARIN. SI-embed.hr (lemma)	100D avg	0.837
	100D	0.927		100D	0.771
w2v Kontekst.io	256D avg	0.483			
	256D	0.718			
	256D lem avg	0.483			
	256D lem	0.718			

Note. The number of all cases is 468 (from 234 occupation pairs) for Slovene and 750 (from 375 occupation pairs) for Croatian.

Appendix B: Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2020

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2020

Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev

Anka Supej*, Matej Ulčar†, Marko Robnik-Šikonja†, Senja Pollak*

*Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
a.supej@gmail.com, senja.pollak@ijs.si

†Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana
{matej.ulcar, marko.robnik}@fri.uni-lj.si

Povzetek

V zadnjih letih je uporaba globokih nevronskih mrež in gostih vektorskih vložitev za predstavitve besedil privedla do vrste odličnih rezultatov na področju računalniškega razumevanja naravnega jezika. Prav tako se je pokazalo, da vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Ena izmed metod za ocenjevanje kakovosti vložitev so izračuni analogij. Prispevek se osredotoča na evalvacijo vektorskih vložitev besed v slovenščini z vidika spola. Sestavili smo seznam moških in ženskih ustreznic poklicev (dostopen prek repozitorija CLARIN) in preko analogij ovrednotili spolno pristranost modelov vložitev fastText, word2vec in ELMo z različnimi konfiguracijami in pristopi k računanju analogij.

Abstract

In recent years, the use of deep neural networks and dense vector embeddings for text representation have led to excellent results in the field of computational understanding of natural language. It has also been shown that word embeddings often capture gender, racial and other types of bias. One of the methods for assessing the quality of word embeddings are analogies calculations. The article focuses on the evaluation of Slovene word embeddings in terms of gender. We compiled a list of male and female equivalents of occupations and evaluated the gender bias of fastText, word2vec and ELMo embeddings.

1. Uvod

Raziskave na stičišču spola in jezika so metodološko različne. Sociolingvistične študije poročajo o načinih, po katerih se uporaba jezika med ženskami in moškimi razlikuje (npr. širše besedišče, milejše izražanje, uporaba tipičnih slovničnih struktur pri ženskah) (Lakoff, 1973; Tannen, 1990; Argamon et al., 2003). Opažanja, da se uporaba jezika med spoloma razlikuje, so navdihnila študije profiliranja avtorjev na besedilih različnih jezikov in tipov besedil (Koolen in van Cranenburgh, 2017; Pardo et al., 2015; Martinc et al., 2017), tudi za slovenščino ((Verhoeven et al., 2017; Škrjanec et al., 2018).

Dimenzija spola v korpusih ni prisotna le kot jezikovna variacija, temveč tudi v obliki večplastne pristranosti, tako v posameznih besedilih kot tudi v večjih korpusih. Sorodne raziskave ugotavljajo:

- da se pristranost odraža kot pomanjkanje omemb žensk: korpusi, ki se pogosto uporabljajo v raziskavah, vsebujejo znatno manj zaimkov ženskega spola (Zhao et al., 2018) ali drugih nanašalnic na ženske (Caldas-Coulhard in Moon, 2010; Baker, 2010);
- da so ženske manj pogosto avtorice ali urednice (Hill in Shaw (2013): le 16% urednic Wikipedije je žensk);
- da korpusi zajemajo spolno stereotipne kolokacije (Pearce, 2008), ki npr. predstavljajo ženske predvsem skozi reproduktivno funkcijo (Gorjanc, 2007) in jih ne povezujejo z (družbeno) močjo (Baker, 2010).

V zadnjih letih je razmah na področju obdelave naravnega jezika povezan predvsem z uporabo globokih nevronskih mrež, ki se uporabljajo tudi za učenje predstavitev besedil v obliki gostih vektorskih vložitev besed. Izkaže se, da tudi vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Pristranost se v besednih vložitvah kaže preko semantičnih asociacij in posledične bližine v vektorskem prostoru (Mikolov et al., 2013b). Računsko jo lahko ovrednotimo npr. s kosinusno podobnostjo med vložitvami, ki opisujejo nek širši pojem (npr. spol), ter stereotipnimi koncepti (npr. v Caliskan et al. (2017): asociacija žensk in umetnosti ter moških in znanosti) ali preko izračuna analogij (Bolukbasi et al., 2016), ki predpostavljajo odnos: $\overrightarrow{moški} - \overrightarrow{poklic_M} \approx \overrightarrow{ženska} - \overrightarrow{poklic_Ž}$. Poleg študij, ki so pokazale na pristranost samih vložitev, so različni avtorji pokazali tudi prenos pristranosti v algoritme za različne naloge obdelave naravnega jezika, od strojnega prevajanja (Prates et al., 2020; Vanmassenhove et al., 2018) do študij sentimenta (Kiritchenko in Mohammad, 2018)). Na drugi strani pa nekateri avtorji (Nissim et al., 2019) opozarjajo tudi na pretirano poudarjanje pristranosti pri zasnovi raziskav z analogijami.

Glavni doprinos prispevka je evalvacija slovenskih modelov besednih vektorskih vložitev z vidika spola, ki še ni dovolj raziskano (izjema je npr. analiza slovenskega modela w2v v Supej et al. (2019)). Prispevek se osredotoča na kvantitativno evalvacijo in primerjavo širokega nabora izbora slovenskih vložitev ter različnih pristopov k evalvaciji, s čimer nagovarja predvsem razvijalce jezikovnotehnoloških orodij, ki vložitve uporabljajo. Kljub temu da s tem indirektno problematiziramo pristranost v jeziku ter pokažemo tudi na nekaj stereotipnih povezav, pa je podrobnejša kritična interpretacija izven fokusa tega prispevka.

V prispevku najprej predstavimo sorodna dela (2. razdelek). V 3. razdelku opišemo seznam moških in ženskih poklicev ter uporabljenih besednih vektorskih vložitev. V 4. in 5. razdelku predstavimo metodologijo in rezultate, nato zaključimo z diskusijo in načrti za nadaljnje delo.

2. Sorodna dela

Korpusi odražajo jezikovne variacije (vključno z različnimi vrstami pristranosti) v odnosu do družbenih dejavnikov. Orodja procesiranja naravnega jezika, ki se učijo na korpusih, lahko variacije in pristranosti podedujejo: nekatere študije prikažejo, da so orodja procesiranja naravnega jezika uspešnejša, ko tovrstne variacije upoštevajo (Volkova et al., 2013; Hovy, 2015). Študija Hovy (2015) pokaže, da vključitev informacij o starosti in spolu avtorjev izboljša uspešnost treh nalog v petih različnih jezikih. Pristranost v korpusih ima lahko tudi negativne posledice, kar lahko podkrepimo z nekaj primeri.

Pogosto uporabljeni korpusi vsebujejo pristranosti do te mere, da so orodja procesiranja naravnega jezika uspešnejša pri vhodnih podatkih, kjer je besedilo napisala starejša oseba (Hovy in Søgaard, 2015). Študija Garimella et al. (2019) pokaže, da sta oblikoslovni označevalnik in skladenjski razčlenjevalnik uspešna na tekstih, ki so jih napisale ženske, ne glede na to, na katerih podatkih sta bila naučena. Teksti moških avtorjev so bolje razčlenjeni/označeni, v kolikor je v učnih podatkih na voljo dovolj besedil, ki so jih napisali moški. Uspešnost orodij, kot so razčlenjevalniki, za tekste moških avtorjev je torej lahko posledica neuravnoteženosti množice učnih podatkov v prid moškemu avtorstvu. Pristranost v korpusih ima poleg negativnega vpliva na orodja obdelave naravnega jezika (Sun et al., 2019) tudi druge negativne posledice, kot je npr. nepravilna razrešitev koreferenc (Zhao et al., 2018).

Vektorske vložitve besed, prav tako naučene na korpusih, poleg sintaktičnih značilnosti besed ujamejo tudi njihove semantične relacije. To se izraža v geometriji prostora vektorskih vložitev: semantično povezane vložitve so si v vektorskem prostoru bližje in razporejene v podobnih smereh. Zato je z njimi mogoče računati tudi odnose, ki presegajo enostavno sorodnost besed, npr. preko analogij. Npr. odnos *Madrid:Španija* je podoben odnosu *Pariz:Francija* (Mikolov et al., 2013b). Vektorske vložitve besed so naučene na korpusih z različnimi algoritmi in tako kot korpusi vsebujejo pristranosti. Beseda, ki je npr. stereotipno povezana z določenim spolom, bo v vektorskem prostoru tako bližje vektorski vložitvi besed *ženska* ali *moški* (Garg et al. (2018) npr. pokažejo, da je pridevnik *časten* v angleščini bližje besedi *moški* kot besedi *ženska*), pristranosti pa se kažejo tudi preko stereotipnih rešitev analogij (npr. Bolukbasi et al. (2016): rešitev analogije *moški:programer::ženska:x* je *gospodinja*). Nissim et al. (2019) opozarjajo, da tovrstne raziskave pretirano poudarjajo pristranost.

Ker se lahko pristranost z uporabo orodij, ki uporabljajo vložitve, ojača (Zhao et al., 2017), se več raziskovalnih skupin ukvarja z metodami "razpristranjevanja" (angl. *debiasing*) vektorskih vložitev. Primeri teh postopkov so izenačevanja oddaljenosti med spolno zaznamovanimi besedami in poklici (Bolukbasi et al., 2016; Bordia in Bowman, 2019), vstavljanje dodatnih omejitev v učni korpus (npr. zagotavljanje enake razporeditve poklicnih aktivnosti med spoloma v učnih podatkih) (Zhao et al., 2017), odstranjevanja tekstov, ki povzročajo pristranost (Brunet et al., 2018), in učenje spolno nevtralnih vektorskih vložitev (Zhao et al., 2018). Gonen in Goldberg (2019) opozarjata, da mnogi postopki "razpristranjevanja" pristranost le zakrijejo, medtem ko ta dejansko ostane prisotna v vložitvah.

Študije na področju raziskovanja pristranosti v vektorskih vložitvah besed so pogosto zasnovane na analogijah poklicev. Ker je v slovenščini spol besed morfološko izražen, kot rezultat analogije pričakujemo žensko oz. moško obliko poklica. Predhodna raziskava na besednih vložitvah (word2vec) v slovenščini (Supej et al., 2019) je pokazala, da je natančnost iskanja analogij dokaj visoka tako pri iskanju moškega kot tudi ženskega poklica. Rezultati kljub temu odražajo spolne pristranosti: rezultat analogije *ženska:tajnica :: moški:x* da rezultat $x =$ *šef*, prvih 10 najbližjih rezultatov različnih analogij pa odseva več spolnih neenakosti: asociacija žensk z domačimi opravili, moških s poklici višjega statusa itd. V nasprotju s predhodno raziskavo se naš prispevek ne ukvarja s sociološko problematizacijo rezultatov analogij enega tipa besednih vložitev (tj. word2vec), temveč preko analogij poklicev ovrednoti različne modele vložitev, njihove konfiguracije in morebiten vpliv filtriranja podatkov na rezultate. V prispevku torej razširimo študijo (Supej et al., 2019) z obširno analizo razpoložljivih modelov slovenskih vektorskih vložitev besed na razširjenem seznamu poklicev.

3. Podatki

V tem razdelku predstavimo sestavljen seznam poklicev ter opišemo različne vektorske vložitve besed.

3.1. Seznam poklicev

Naš izbor poklicev temelji na standardni klasifikaciji poklicev (Vlada RS, 1997), katere osnova je *Mednarodna standardna klasifikacija poklicev*. Večina poklicev v klasifikaciji je večbesednih zvez (npr. *upravljalec/upravljalka metalurškega žerjava*), ki so zaradi svoje specifičnosti in obsežnosti manj primerne za računske naloge. Za potrebe izračuna analogij smo se omejili na enobesedne poklice. Celotni seznam enobesednih poklicev zajema 422 parov, ki jih omejimo še glede na naslednje kriterije:

(1) Poklic ima žensko in moško obliko (spolno nevtralne besede, npr. *pismonoša*, niso vključene). (2) Vsaj ena izmed oblik poklica se pojavi v Slovenskem oblikoslovnem leksikonu *Sloleks 2.0* (Dobrovoljc et al., 2019), ki omogoča razlikovanje med lastnimi in občnimi imeni (nekateri poklici so namreč tudi lastna imena; npr. *kovač*), ali pa se v referenčnem korpusu standardne slovenščine Gigafida 2.0 (2020) pojavi 500- ali večkrat. (3) V primerih, kjer ocenimo, da za poimenovanje v standardni klasifikaciji poklicev obstajajo bolj uveljavljene različice, naboru podatkov dodamo sopomenko z istim korenom (npr. za izraz *fotografka*) iz standardne klasifikacije dodamo sopomenko *fotografinja*). Pri izračunih za izhodiščne besede upoštevamo obliko, ki je v korpusih bolj pogosta, pri pravilnosti izračunanih analogij pa upoštevamo katerokoli različico. (4) Če standardna klasifikacija ne vključuje

ženske (npr. *dramatik*) ali moške oblike poklica (npr. *prostitutka*), smo ročno dodali ustrezno različico, v kolikor ta obstaja (npr. za *postreščka* in za *hosteso* uveljavljene ženske oz. moške oblike ni) in je prisotna v Gigafidi. (5) Iz nabora smo izključili poklice, kjer je ženska oz. moška varianta poklica homofon (npr. *strežnik*, *detektivka*), oz. kjer je poklic možno asociirati s poklici nepovezanim kontekstom (npr. *čarovnik/čarovnica*).

Končni seznam vsebuje 234 parov poklicev, ki bo prosto dostopen na repozitoriju CLARIN [1].

3.2. Modeli vektorskih vložitev

V eksperimentih smo uporabili več različnih konfiguracij znanih vektorskih vložitev:

- fastText (Bojanowski et al., 2016):
  - 100-dimenzionalni vektorji, naučeni tekom projekta EMBEDDIA[2] na Gigafidi 2.0,
  - 300-dimenzionalni vektorji, naučeni kot v prejšnjem primeru,
  - 100-dimenzionalni vektorji besed s portala Sketch Engine (word),
  - 100-dimenzionalni vektorji s portala Sketch Engine, kjer so vektorji vložitve lem (lemma),
  - 100-dimenzionalni vektorji CLARIN.SI-embed.sl (Ljubešić in Erjavec, 2018) in
  - 300-dimenzionalni vektorji s portala fasttext.cc;
- word2vec (Mikolov et al., 2013a): 256-dimenzionalni vektorji, ki so bili naučeni za potrebe portala Kontekst.io (Plahuta, 2020) in so na voljo po dogovoru[3];
- ELMo (Peters et al., 2018): 1024-dimenzionalni vektorji kontekstnih vložitev projekta *EMBEDDIA*, naučeni na Gigafidi (Ulčar, 2019), kjer so vzete povprečne vrednosti 200.000 najpogostejših besed (izračunano na podlagi slovenske Wikipedije). Uporabljenih je bilo več različnih vrst vektorjev:
  - vektorji z izhoda prvega (CNN) nivoja mreže, ki so kontekstno neodvisni (tj. *layer 0*),
  - vektorji z izhoda drugega (prvega LSTM) nivoja mreže, ki so kontekstno odvisni (tj. *layer 1*),
  - vektorji z izhoda tretjega (drugega LSTM) nivoja mreže, ki so kontekstno odvisni (tj. *layer 2*).

4. Metodologija evalvacije

Analogije poklicev smo za vsako izmed vložitev izračunali na štiri različne načine. Jedro pristopa je pri vseh načinih enako: za vsako moško obliko poklica (P_m) iščemo ustrezno žensko obliko (P_f). Izračunamo vektor

$$\overrightarrow{d} = \overrightarrow{P_m} - \overrightarrow{m} + \overrightarrow{f},$$

kjer je $\overrightarrow{m}$ moški vektor in $\overrightarrow{f}$ ženski vektor. V idealnem primeru bi bil vektor $\overrightarrow{d}$ enak $\overrightarrow{P_f}$. Vektorju $\overrightarrow{d}$ poiščemo vektorje N najbližjih besed glede na kosinusno razdaljo.

[1] http://hdl.handle.net/11356/1347
[2] http://embeddia.eu/
[3] https://kontekst.io/partnerstvo

m	f	m	f
moški	ženska	brat	sestra
gospod	gospa	oče	mati
fant	dekle	sin	hči
fant	punca	dedek	babica
deček	deklica	mož	žena
stric	teta	on	ona

Tabela 1: Pari inherentno moških in ženskih besed.

Pri iskanju najbližjih besed smo upoštevali vse besede, ki se nahajajo v vložitvah, razen besed *moški*, *ženska*, besede P_m ter vseh besed, ki vsebujejo nečrkovne simbole (številke, vezaje, druga ločila, itd.) Če se beseda P_f nahaja med N najbližjimi besedami, ta primer štejemo kot pravilno določenega, sicer kot napačnega. Pri tem smo ignorirali velike in male začetnice, na primer besede *Zdravnik*, *zdravnik* in *ZDRAVNIK* upoštevamo kot isto besedo.

Postopek ponovimo za vsako žensko obliko poklica (P_f), kjer iščemo ustrezno moško obliko (P_m). Vektor $\overrightarrow{d}$ v tem primeru izračunamo kot

$$\overrightarrow{d} = \overrightarrow{P_f} - \overrightarrow{f} + \overrightarrow{m},$$

pri iskanju najbližjih besed pa namesto besede P_m izpustimo besedo P_f. Končni rezultat predstavlja delež pravilno določenih primerov, oz. mera *natančnost pri* N (angl. *precision at* N oz. $P@N$).Višji N omogoča primerjavo v širši okolici zadetka v vektorskem prostoru.

Za določitev moškega vektorja $\overrightarrow{m}$ in ženskega vektorja $\overrightarrow{f}$ smo uporabili dva pristopa. V prvem je m kar beseda *moški* in f beseda *ženska*. V drugem pristopu razliko $\overrightarrow{f} - \overrightarrow{m}$, oz. $\overrightarrow{m} - \overrightarrow{f}$ podobno kot Bolukbasi et al. (2016) predstavimo s povprečno razliko vektorjev parov besed, ki se specifično nanašajo na žensko oz. moškega (Tabela 1).

Pri iskanju najbližjih N besed smo uporabili tudi alternativen pristop, kjer smo vse besede v vložitvah lematizirali z orodjem LemmaGen[4]. S tem smo izničili vpliv pregibanja besed; na primer, besedi *zdravnico* in *zdravnice* sta pri tem postopku enaki, saj imata isto lemo *zdravnica*.

5. Rezultati

Rezultate predstavimo za vsak pristop, opisan v 4. razdelku, z mero natančnost pri N, kjer je N enak 1, 5 in 10. Nekaterih poklicev z našega seznama ni v vseh vložitvah. Če iskane besede ni med N najbližjimi, je primer označen kot napačen, četudi te besede sploh ni med vložitvami. Primere, ko poklica, ki ga imamo na vhodu, ni med vložitvami in tako ne moremo izračunati vektorja $\overrightarrow{d}$, obravnavamo na dva načina. V prvem načinu (`all`) tak primer štejemo kot napačen, v drugem načinu (`covered`) pa ga izločimo iz primerov in na končni rezultat ne vpliva.

Rezultati, kjer imamo na vhodu moški poklic P_m in iščemo ustrezni ženski poklic P_f so v Tabeli 2. Rezultati, kjer za ženski poklic P_f na vhodu iščemo moški poklic P_m so v Tabeli 3. Pristop, pri katerem smo vse besede lematizirali, ima pripono `_lem`. Pristop, kjer smo za moški in ženski vektor oz. njuno razliko uporabili povprečne razlike vektorjev besed iz tabele 1, ima pripono `_avg`.

[4] https://github.com/vpodpecan/lemmagen3/

Vložitve	št. dimenzij in pristop	all P@1	all P@5	all P@10	covered P@1	covered P@5	covered P@10
ELMo Embeddia	1024D_l0_avg	0.128	0.278	0.299	0.166	0.359	0.387
	1024D_l0	0.162	0.291	0.308	0.210	0.376	0.398
	1024D_l0_lem_avg	0.286	0.308	0.312	0.370	0.398	0.403
	1024D_l0_lem	0.291	0.303	0.312	0.376	0.392	0.403
	1024D_l1_avg	0.295	0.303	0.308	0.381	0.392	0.398
	1024D_l1	0.291	0.303	0.303	0.376	0.392	0.392
	1024D_l1_lem_avg	0.295	0.303	0.308	0.381	0.392	0.398
	1024D_l1_lem	0.291	0.303	0.303	0.376	0.392	0.392
	1024D_l2_avg	0.291	0.308	0.308	0.376	0.398	0.398
	1024D_l2	0.286	0.308	0.308	0.370	0.398	0.398
	1024D_l2_lem_avg	0.291	0.308	0.308	0.376	0.398	0.398
	1024D_l2_lem	0.286	0.308	0.308	0.370	0.398	0.398
fastText.cc	300D_avg	0.594	0.722	0.735	0.607	0.738	0.751
	300D	0.436	0.688	0.718	0.445	0.703	0.734
	300D_lem_avg	0.641	0.739	0.748	0.655	0.755	0.764
	300D_lem	0.487	0.709	0.735	0.498	0.725	0.751
fastText Embeddia	100D_avg	0.667	0.709	0.714	0.672	0.716	0.720
	100D	0.632	0.709	0.714	0.638	0.716	0.720
	100D_lem_avg	0.671	0.714	0.718	0.677	0.720	0.724
	100D_lem	0.632	0.709	0.714	0.638	0.716	0.720
	300D_avg	0.662	0.709	0.718	0.668	0.716	0.724
	300D	0.679	0.714	0.714	0.685	0.720	0.720
	300D_lem_avg	0.679	0.714	0.718	0.685	0.720	0.724
	300D_lem	0.679	0.714	0.714	0.685	0.720	0.720
fastText CLARIN.SI-embed.sl	100D_avg	**0.761**	0.868	0.880	**0.761**	0.868	0.880
	100D	0.705	0.855	0.885	0.705	0.855	0.885
	100D_lem_avg	**0.761**	**0.880**	**0.902**	**0.761**	**0.880**	**0.902**
	100D_lem	0.709	0.859	0.885	0.709	0.859	0.885
fastText Sketch Engine (word)	100D_avg	0.714	0.765	0.774	0.717	0.768	0.777
	100D	0.688	0.765	0.774	0.691	0.768	0.777
	100D_lem_avg	0.722	0.778	0.782	0.725	0.781	0.785
	100D_lem	0.688	0.765	0.778	0.691	0.768	0.781
fastText Sketch Engine (lemma)	100D_avg	0.598	0.786	0.821	0.598	0.786	0.821
	100D	0.380	0.658	0.756	0.380	0.658	0.756
word2vec Kontekst.io	256D_avg	0.402	0.543	0.585	0.407	0.550	0.593
	256D	0.248	0.483	0.509	0.251	0.489	0.515
	256D_lem_avg	0.402	0.543	0.585	0.407	0.550	0.593
	256D_lem	0.248	0.483	0.513	0.251	0.489	0.519

Tabela 2: Rezultati za vse vložitve in variante, kjer je na vhodu moški poklic in iščemo ustrezen ženski poklic. Če za moški poklic na vhodu ne najdemo vložitve, tak primer štejemo kot napačno ugotovljen (all), oz. ga izpustimo iz rezultatov (covered). Najboljši rezultati v vsakem stolpcu so odebeljeni.

Rezultati kažejo, da dobimo boljše rezultate s fastText vložitvami, z izračunom, kjer namesto samega vektorja *moški* oz. *ženska* uporabimo povprečje besed z inherentno izraženim spolom ter z lematizacijo. Rezultate podrobneje razčlenimo v naslednji sekciji.

6. Diskusija

Vložitve, ki dosegajo največjo natančnost pri ugotavljanju analogij (z vhodnim moškim poklicem), so vložitve fastText CLARIN.SI-embed.sl (Tabela 2). Pri vhodnem ženskem poklicu dosegajo največjo natančnost, če upoštevamo le vložitve poklicev, ki so prisotni, fastText Embeddia, medtem ko so na vzorcu vseh vložitev najnatančnejše vložitve fastText CLARIN.SI-embed.sl (Tabela 3). V različnih modelih vložitev, pri različnih vhodnih podatkih velja, da lematizacija izhodnih podatkov in hkrati uporaba vektorja povprečne razlike med ženskimi in moškimi besedami (namesto uporabe le besed *ženska* oz. *moški*) izboljša natančnost analogije. Modeli, kjer je na vhodu poklic ženskega spola, v povprečju dosegajo višjo natančnost analogij v primeru `covered` rezultatov (če ženskega poklica ni v med vložitvami, tega ne štejemo kot napačno ugotovljeno analogijo). Rezultati `all` so podobni pri obeh tipih vhodnih podatkov.

Vložitve fastText Embeddia dosegajo zelo podobne rezultate s 100- in 300-dimenzionalnimi vložitvami, (glej Tabeli 2 in 3). (Druge vložitve so bile naučene na drugih jezikovnih virih, zato niso neposredno primerljive.) Vendar pa je iz Tabele 5 (vložitve fastText Embeddia) tudi razvidno, da igra dimenzionalnost veliko vlogo pri tem, kako pogosto je rezultat analogije sam vhodni poklic. Dimenzionalnost bi torej imela velik vpliv na nefiltrirane rezultate.

V vseh modelih vložitev je delež poklicev moškega spola večji kot delež poklicev ženskega spola (Tabela 4).

Vložitve	št. dimenzij in pristop	all P@1	all P@5	all P@10	covered P@1	covered P@5	covered P@10
ELMo Embeddia	1024D_l0_avg	0.226	0.299	0.303	0.707	0.933	0.947
	1024D_l0	0.137	0.295	0.303	0.427	0.920	0.947
	1024D_l0_lem_avg	0.291	0.299	0.303	0.907	0.933	0.947
	1024D_l0_lem	0.286	0.303	0.303	0.893	0.947	0.947
	1024D_l1_avg	0.291	0.303	0.303	0.907	0.947	0.947
	1024D_l1	0.282	0.303	0.303	0.880	0.947	0.947
	1024D_l1_lem_avg	0.291	0.303	0.303	0.907	0.947	0.947
	1024D_l1_lem	0.291	0.303	0.303	0.907	0.947	0.947
	1024D_l2_avg	0.282	0.299	0.299	0.880	0.933	0.933
	1024D_l2	0.274	0.295	0.299	0.853	0.920	0.933
	1024D_l2_lem_avg	0.282	0.299	0.299	0.880	0.933	0.933
	1024D_l2_lem	0.274	0.295	0.299	0.853	0.920	0.933
fastText.cc	300D_avg	0.291	0.590	0.675	0.393	0.798	0.913
	300D	0.111	0.415	0.585	0.150	0.561	0.792
	300D_lem_avg	0.453	0.654	0.701	0.613	0.884	0.948
	300D_lem	0.338	0.637	0.679	0.457	0.861	0.919
fastText Embeddia	100D_avg	0.654	0.705	0.709	0.900	0.971	0.976
	100D	0.342	0.632	0.658	0.471	0.871	0.906
	100D_lem_avg	0.658	0.705	0.709	0.906	0.971	0.976
	100D_lem	0.534	0.671	0.684	0.735	0.924	0.941
	300D_avg	0.607	0.705	0.709	0.835	0.971	0.976
	300D	0.239	0.624	0.697	0.329	0.859	0.959
	300D_lem_avg	0.688	0.709	0.714	**0.947**	**0.976**	**0.982**
	300D_lem	0.594	0.705	0.709	0.818	0.971	0.976
fastText CLARIN.SI-embed.sl	100D_avg	0.731	0.850	0.876	0.784	0.913	0.940
	100D	0.077	0.547	0.726	0.083	0.587	0.780
	100D_lem_avg	**0.782**	**0.876**	**0.885**	0.839	0.940	0.950
	100D_lem	0.607	0.821	0.855	0.651	0.881	0.917
fastText Sketch Engine (word)	100D_avg	0.701	0.761	0.769	0.886	0.962	0.973
	100D	0.167	0.598	0.718	0.211	0.757	0.908
	100D_lem_avg	0.735	0.761	0.769	0.930	0.962	0.973
	100D_lem	0.641	0.752	0.761	0.811	0.951	0.962
fastText Sketch Engine (lemma)	100D_avg	0.581	0.803	0.829	0.673	0.931	0.960
	100D	0.440	0.701	0.769	0.510	0.812	0.891
word2vec Kontekst.io	256D_avg	0.453	0.568	0.581	0.679	0.853	0.872
	256D	0.244	0.393	0.479	0.365	0.590	0.718
	256D_lem_avg	0.453	0.568	0.581	0.679	0.853	0.872
	256D_lem	0.342	0.457	0.530	0.513	0.686	0.795

Tabela 3: Rezultati za vse vložitve in variante, kjer je na vhodu ženski poklic in iščemo ustrezen moški poklic. Če za ženski poklic na vhodu ne najdemo vložitve, tak primer štejemo kot napačno ugotovljen (all), oz. ga izpustimo iz rezultatov (covered). Najboljši rezultati v vsakem stolpcu so odebeljeni.

Največja pokritost je v vložitvah fastText CLARIN.SI-embed.sl, pri vložitvah modela ELMo pa se na primer pojavi le 75 od 234 izbranih poklicev ženskega spola. Razlog za mnogo manjšo zastopanost poklicev pri modelu ELMo je, da smo se zaradi tehničnih razlogov omejili le na 200 tisoč najpogostejših besed v Wikipediji (ELMo vložitve so v osnovi kontekstualne vložitve in je proces povprečenja računsko zahteven). Pri drugih tehnologijah vložitev smo imeli približno milijon besed. Moški poklici, ki se ne pojavljajo v vložitvah, so običajno poklici, ki so tipično povezani z ženskim spolom (npr. *šiviljec* ali *kozmetik*). Tudi poklici ženskega spola, ki se ne pojavljajo v vložitvah, so npr. tradicionalno povezani z moškimi (v vložitvah različnih tipov na primer ni *avtomehaničarke, tesarke* itd.) ali pa gre za kulturno pogojene izključno moške poklice (npr. *nadškof*). Slabo zastopanost poklicev ženskega spola lahko povežemo tudi z drugimi faktorji – Zhao et al. (2018) poročajo, da se različne zvrsti tekstov pogosteje nanašajo na moške v okviru njihovega poklica kot pa je to pri ženskah.

Modeli vložitev v konfiguraciji `lem_avg` (uporaba vektorja povprečnih razlik med spolno zaznamovanimi besedami in lematiziranje izhodnih podatkov) dajejo zelo različne rezultate. Rezultati analogij pri modelih ELMo in word2vec so večinoma poklici. Pri vložitvah fastText Embeddia, CLARIN.SI-embed.sl in Sketch Engine (word) so rezultati poklici in ostale besede, sorodne vhodnemu poklicu, ter besede z istim korenom kot vhodni poklic. Rezultati modelov fastText.cc in Sketch Engine (lemma) so večinoma besede z istim korenom kot vhodni poklic.

Po mnenju Nissim et al. (2019) je interpretacija večine študij, ki povezujejo analogije s pristranostjo, pretirana. Računanje analogij je namreč zastavljeno tako, da se izključi vhodni poklic, četudi bi bil to dejanski rezultat z

najvišjo kosinusno podobnostjo. Kljub temu, da smo pri rezultatih izločili vhodne poklice, kar je za samo računanje analogij standarden postopek, smo analizirali tudi rezultate pred filtriranjem. Pri analizi teh rezultatov smo opazili, da je med rezultati analogije z najvišjo kosinusno podobnostjo pogosto sam vhodni poklic (Tabela 5), kar pa zelo variira med posameznimi modeli.

Rezultati analogij so zanimivi z vidika semantike. Prva rezultata analogij (vložitev fT Embeddia 100D_lem_avg) *ženska:krojačica :: moški:x krojač* in *ženska:šivilja :: moški:x krojač* ponazarjata, da besedne vektorske vložitve upoštevajo tako slovnične kot tudi semantične elemente (vektorske vložitve besede *šiviljec* ni, *krojač* pa je semantično povezana beseda). Rešitve nekaterih analogij (predvsem v modelu w2v Kontekst.io lem_avg) z vhodnim poklicem niso povezane ali so stereotipne. Npr., rešitve analogije *moški:rudar :: ženska:x* v modelu w2v Kontekst.io lem_avg so npr.: *barbika, klovnesa, čarovnica, lutka, prostitutka, akrobatka, najstnica, opica, princeska, striptizeta*. Na stereotipne analogije v modelu w2v opozorijo tudi v Supej et al. (2019).

V okviru analize smo naredili tudi skupni frekvenčni seznam rezultatov analogij za vse ženske oz. moške vhodne poklice za posamezen model vložitev (upoštevajoč le konfiguracije `lem_avg`) (cf. Tabela 6). Opazimo vzorec, da se pri modelih ELMo l2_lem_avg in w2v Kontekst.io lem_avg najbolj pogosti ženski poklici/besede pojavljajo pogosteje kot najpogostejši moški poklici. Ena od možnih interpretacij je, da izbrana modela v primerjavi z nekaterimi drugimi vsebujeta relativno manj vektorskih besednih vložitev (200.000 oz. približno 600.000 za posamezen model). Oba modela imata tudi manjšo zastopanost ženskih oblik poklicev med besednimi vložitvami. Poklici, ki se kljub temu pojavljajo med vložitvami, se zato ponovijo večkrat. Poklicev moškega spola je v besednih vložitvah več, zato se posamezni poklici ne pojavljajo tako pogosto.

Med pogostimi ženskimi analogijami pri modelih ELMo l2_lem_avg in w2v Kontekst.io lem_avg zaznamo poklice nižjega družbenega statusa (*čistilka, perica, gospodinja*) ter zastarele poklice, kjer je bila ženska v podrejenem položaju (*služkinja*). Pri najpogostejših moških analogijah so poklici nižjega družbenega statusa izjemno redki.

Ugotavljamo tudi, da se nekatere besede (predvsem poklici ženskega spola) v rezultatih pojavljajo ne glede na semantično povezanost z vhodnim poklicem. V več primerih je rešitev analogije (predvsem ko gre za vhodni tipično moški poklic) nepovezana z vhodnim poklicem (npr. *bolničarka* kot prva rešitev analogije *moški:rudar :: ženska:x* in *šivilja* kot prva rešitev analogije *moški:avtomehanik :: ženska:x* v modelu fT Embeddia 100D_lem_avg). Možna razlaga, za potrditev katere bi bili potrebni dodatni testi, je, da so nekatere vektorske vložitve besed bolj 'centralne' od drugih in so najbližji sosed velikemu številu drugih besed, kar je v vektorskih vložitvah mogoč pojav. Možnost za nadaljnje delo je (delno) zmanjšati vpliv tovrstnih vložitev s pomočjo mere, alternativne kosinusni podobnosti, tj. CSLS (Conneau et al., 2018) oz. podobne mere, ki upošteva medsebojne razdalje najbližjih *n* sosedov).

vložitve	m	f
ELMo	0.774	0.321
fastText_cc	0.979	0.739
fastText Embeddia	0.991	0.726
fastText CLARIN.SI-embedd.sl	**1.000**	**0.932**
fastText Sketch Engine (word)	0.996	0.791
fastText Sketch Engine (lemma)	1.000	0.863
word2vec Kontekst.io	0.987	0.667

Tabela 4: Delež poklicev moškega (m) in ženskega (f) spola v vložitvah.

Vložitve	št. dim. in pristop	delež
ELMo Embeddia	1024D_l0_avg	0.547
	1024D_l0	0.547
	1024D_l0_lem_avg	0.547
	1024D_l0_lem	0.547
	1024D_l1_avg	0.423
	1024D_l1	0.483
	1024D_l1_lem_avg	0.423
	1024D_l1_lem	0.483
	1024D_l2_avg	0.064
	1024D_l2	0.088
	1024D_l2_lem_avg	0.064
	1024D_l2_lem	0.088
fT fastText.cc	300D_avg	0.831
	300D	0.825
	300D_lem_avg	0.831
	300D_lem	0.825
fT Embeddia	100D_avg	0.143
	100D	0.141
	100D_lem_avg	0.143
	100D_lem	0.141
	300D_avg	0.419
	300D	0.513
	300D_lem_avg	0.419
	300D_lem	0.513
fT CLARIN.SI-embed.sl	100D_avg	0.316
	100D	0.310
	100D_lem_avg	0.316
	100D_lem	0.310
fT Sketch Engine (word)	100D_avg	0.096
	100D	0.135
	100D_lem_avg	0.096
	100D_lem	0.135
fT Sketch Engine (lemma)	100D_avg	0.803
	100D	0.927
w2v Kontekst.io	256D_avg	0.483
	256D	0.718
	256D_lem_avg	0.483
	256D_lem	0.718

Tabela 5: Delež primerov, pri katerih je rezultat analogije z najvišjo kosinusno podobnostjo sam vhodni poklic (pred filtriranjem za računanje rezultatov v Tabelah 2 in 3). Št. vseh primerov je 468 iz 234 parov poklicev.

7. Zaključki in nadaljnje delo

V prispevku smo na nalogi analogij moških in ženskih poklicev ovrednotili različne slovenske vektorske vložitve (z različnimi konfiguracijami in pristopi k računanju analogij). Ugotovili smo, da dobimo najboljše rezultate s fastText vložitvami. Pri ženskih analogijah za moške poklice

ELMo Embeddia l2_lem_avg				fastText CLARIN.SI_lem_avg				word2vec Kontekst.io_lem_avg			
m vhod		f vhod		m vhod		f vhod		m vhod		f vhod	
Rezultat	n	Rezultat	n	Rezultat	n	Rezultat	n	Rezultat	n	Rezultat	n
bolničarka	47	geograf	9	šivilja	15	mizar	11	kuharica	44	ortoped	14
biokemičarka	39	politolog	8	ključavničarka	11	biolog	10	gospodinja	38	pisatelj	14
frizerka	39	biolog	7	inštalaterka	9	ključavničar	9	šivilja	33	kardiolog	13
trgovka	39	dramaturg	7	keramičarka	9	zgodovinar	9	frizerka	32	nevrolog	13
čistilka	34	književnik	7	filologinja	8	internist	8	kozmetičarka	30	urolog	13
znanstvenica	34	scenarist	7	oftalmologinja	8	režiser	8	čistilka	29	psihiater	12
kuharica	33	animator	6	filozofinja	7	arheolog	7	fotografinja	29	ekolog	11
geologinja	30	esejist	6	geofizičarka	7	natakar	7	zdravnica	29	hišnik	11
perica	28	etnolog	6	kmetica	7	pisatelj	7	služkinja	26	biolog	10
služkinja	28	fotograf	6	nevrokirurginja	7	primarij	7	trgovka	26	korenjak	10
biologinja	27	ilustrator	6	strugarka	7	stomatolog	7	slikarka	25	maneken	10
gospodinja	26	lutkar	6	geologinja	6	tesar	7	tajnica	25	režiser	10
matematičarka	26	paleontolog	6	hematologinja	6	fotoreporter	6	veterinarka	25	akademik	9
mikrobiologinja	26	pravnik	6	kardiologinja	6	gostilničar	6	znanstvenica	25	akademski_slikar	9
arheologinja	25	režiser	6	paleontologinja	6	kardiolog	6	socialna_delavka	24	glasbenik	9

Tabela 6: 15 najpogostejših besed, ki se pojavljajo med prvimi desetimi rezultati analogij v določenem modelu vložitev glede na vhodni poklic v analogiji (*m* ali *f*).

na vhodu se najbolje odreže model fastText CLARIN.SI-embed.sl, za ženske poklice na vhodu pa so to modeli fastText CLARIN.SI-embed.sl ter fastText Embeddia. Pristop, kjer za izračun namesto samega vektorja *moški* oz. *ženska* uporabimo povprečje besed z inherentno izraženim spolom, izboljša rezultate, enako velja za lematizacijo. Najboljši rezultati (P@10) so tako 0.885 za ženske iztočnice z modelom fastText CLARIN.SI-embed.sl-100D_lem_avg in 0.982 s fastText Embeddia 300D_lem_avg z upoštevanim pogojem, da so poklici v vložitvah. Za moške poklice na vhodu pa 0.902 z modelom fastText CLARIN.SI-embed.sl 100D_lem_avg (enak rezultat velja tudi za pogoj prisotnih vložitev). Modeli fastText CLARIN.SI-embed.sl imajo največji delež iskanih moških in ženskih poklicev. Med obravnavanimi vložitvami kažejo vložitve modela Kontekst.io pri kvalitativni analizi na največjo pristranost modela glede na spol (stereotipno ženski in moški poklici, ki se pojavljajo med analogijami ne glede na iztočnico). Prispevek se sicer osredotoča na kvantitativno evalvacijo in je s tem uporaben predvsem za razvijalce novih orodij, podrobnejši kvalitativni analizi in odnosu med vložitevami, jezikom in družbeno močjo pa se bomo posvetili v prihodnje. V nadaljnjem delu bomo obravnavali tudi kontekstne vložitve modela BERT, preizkusili metode za zmanjševanje vpliva vložitev, ki so bolj centralne od drugih, ter študijo razširili na druge jezike projekta EMBEDDIA. Poleg tega bomo preizkusili vpliv spolnih pristranosti v napovednih modelih na praktičnih nalogah, kot je analiza sentimenta.

8. Zahvala

Delo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programov P2-0103 (Tehnologije znanja) in P6-411 (Jezikovni viri in tehnologije za slovenski jezik) ter EU prek okvirnega programa za raziskave in inovacije Obzorje2020 - projekt EMBEDDIA (št. 825153).

9. Literatura

Gigafida 2.0. 2020. Gigafida 2.0: Korpus pisne standardne slovenščine. https://viri.cjvt.si/gigafida, 1. 5. 2020.

Shlomo Argamon, Moshe Koppel, Jonathan Fine in Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346.

Paul Baker. 2010. Will Ms ever be as frequent as Mr? a corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender & Language*, 4(1):125–149.

Piotr Bojanowski, Edouard Grave, Armand Joulin in Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama in Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. V: *30th Conference on Neural Information Processing Systems*.

Shikha Bordia in Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *CoRR*, abs/1904.03035.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson in Richard S. Zemel. 2018. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611.

Carmen Rosa Caldas-Coulhard in Rosamund Moon. 2010. 'curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, 21(2):99–133.

Aylin Caliskan, Joanna J. Bryson in Arvind Narayanan. 2017. Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356(6334):183–186.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer in Hervé Jégou. 2018. Word translation without parallel data. V: *Proc. of International Conference on Learning Representation (ICLR)*.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. Morphological lexicon Sloleks 2.0. CLARIN.SI. http://hdl.handle.net/11356/1230.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky in James Zou. 2018. Word embeddings quantify 100 years of gen-

der and ethnic stereotypes. *PNAS*, 115(16).

Aparna Garimella, Carmen Banea, Dirk Hovy in Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. V: *Proc. of the 57th Annual Meeting of the ACL*, str. 3493–3498. ACL.

Hila Gonen in Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. V: *Proc. of NAACL-HLT 2019*, str. 609–614.

Vojko Gorjanc. 2007. Kontekstualizacija oseb ženskega in moškega spola v slovenskih tiskanih medijih. V: I. Novak-Popov, ur., *Stereotipi v slovenskem jeziku, literaturi in kulturi: zbornik predavanj 43. seminarja slovenskega jezika, literature in kulture*, str. 173–180. Center za slovenščino kot drugi/tuji jezik, Ljubljana.

Benjamin Hill in Aaron Shaw. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS One*, 8.

Dirk Hovy in Anders Søgaard. 2015. Tagging performance correlates with author age. V: *Proc. of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, str. 483–488.

Dirk Hovy. 2015. Demographic factors improve classification performance. V: *Proc. of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, str. 752–762. ACL.

Svetlana Kiritchenko in Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

Corina Koolen in Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. V: *Proc. of the First Ethics in NLP workshop*, str. 12–22. ACL.

Robin Lakoff. 1973. Language and woman's place. *Language in Society*, 2(1):45–80.

Nikola Ljubešić in Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1204.

Matej Martinc, Iza Škrjanec, Katja Zupan in Senja Pollak. 2017. PAN 2017: Author profiling - gender and language variety prediction. V: *Working Notes of CLEF 2017*.

Tomas Mikolov, Greg S. Corrado, Kai Chen in Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. V: *International Conference on Learning Representations*, str. 1–12.

Tomas Mikolov, Wen-tau Yih in Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. V: *Proc. of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies*, str. 746–751. ACL.

Malvina Nissim, Rik Noord in Rob van der Goot. 2019. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, str. 1–17.

Francisco M. Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein in Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. V: L. Cappellato, N. Ferro, G. J. F. Jones in E. SanJuan, ur., *Working Notes of CLEF 2015*, zvezek 1391 iz *CEUR Workshop Proceedings*. CEUR-WS.org.

Michael Pearce. 2008. Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, 3:1–29, 05.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee in Luke Zettlemoyer. 2018. Deep contextualized word representations. V: *Proc. of NAACL-HLT 2018*, str. 2227–2237.

Marko Plahuta. 2020. O slovarju. https://kontekst.io/o-slovarju.

Marcelo O. R. Prates, Pedro H. Avelar in Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang in William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. V: *Proc. of the 57th Annual Meeting of the ACL*, str. 1630–1640. ACL.

Anka Supej, Marko Plahuta, Matthew Purver, Michael Mathioudakis in Senja Pollak. 2019. Gender, language, and society: Word embeddings as a reflection of social inequalities in linguistic corpora. V: *Zbornik Slovenskega sociološkega srečanja 2019 - Znanost in družbe prihodnosti*, str. 75–83.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. Ballantine Books, NY.

Matej Ulčar. 2019. ELMo embeddings model, Slovenian. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1257.

Eva Vanmassenhove, Christian Hardmeier in Andy Way. 2018. Getting gender right in neural machine translation. V: *Proc. of the EMNLP*, str. 3003–3008. ACL.

Ben Verhoeven, Iza Škrjanec in Senja Pollak. 2017. Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. V: *Proc. of the 6th BSNLP Workshop*, str. 119–125. ACL.

Vlada RS. 1997. 1641. uredba o uvedbi in uporabi standardne klasifikacije poklicev. *Uradni list RS*, 28:2217. https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=199728&stevilka=1641.

Svitlana Volkova, Theresa Wilson in David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. V: *Proc. of the EMNLP*, str. 1815–1827. ACL.

Iza Škrjanec, Nada Lavrač in Senja Pollak. 2018. Napovedovanje spola slovenskih blogerk in blogerjev. V: D. Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine*, str. 356–373. Ljubljana: Znanstvena založba FF.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez in Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. V: *Proc. of the EMNLP*, str. 2979–2989. ACL.

Tianlu Zhao, Jieyu fand Wang, Mark Yatskar, Vicente Ordonez in Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. V: *Proc. of the NAACL-HLT*, str. 15–20. ACL.

Appendix C: Gender, Language, and Society - Word Embeddings as a Reflection of Social Inequalities in Linguistic Corpora

ZNANOST IN DRUŽBE PRIHODNOSTI

ANKA SUPEJ
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana
MARKO PLAHUTA
podjetje Virostatik, Ljubljana
MATTHEW PURVER
Queen Mary University of London, London, UK
MICHAEL MATHIOUDAKIS
University of Helsinki, Helsinki, Finland
SENJA POLLAK
Jožef Stefan Institute, Ljubljana

GENDER, LANGUAGE, AND SOCIETY – WORD EMBEDDINGS AS A REFLECTION OF SOCIAL INEQUALITIES IN LINGUISTIC CORPORA

Abstract: *Research on language and gender has a long tradition, and large electronic text corpora and novel computational methods for representing word meaning have recently opened new directions. We explain how gender can be analysed using word embeddings: vector representations of words computationally derived from lexical context in large corpora and capturing a degree of semantics. Being derived from naturally-occurring text, these also capture human biases, stereotypes and reflect social inequalities. The relation between the English words man and programmer can correspond to that between woman and homemaker. In Slovene, the availability of male and female forms for many words for occupations means that such effects might be reduced; however, we study a range of such relations and show that some gender bias still persists (e.g. the relation between words woman and secretary is very similar to that between man and boss).*

Key words: *gender bias, word embeddings, occupations, language and society, natural language processing*

Introduction

Researchers have long been interested in the relationship between language and gender. What started as introspective research into how women and men are discussed and how their way of talking differs (Lakoff 1973), developed into sociolinguistic modelling of discourse styles and different kinds of statistical analyses, which, for example, explore words with which men or women are described. These approaches are now being increasingly complemented by advanced natural language processing (NLP) methods[1], among them *word embeddings* (see below), which can convey meaningful relationships between gender and language.

1. NLP methods are computational methods, designed to process and analyse large amounts of human (i.e. natural) language.

Language can be also understood as being one of the most powerful means through which sexism and gender discrimination are perpetrated and reproduced, via, for example, the content of gender stereotypes, as well as the language structures used (Menegatti and Rubini 2017). The stereotypes reproduced in the lexical choices of everyday communication are not neutral: they reflect the asymmetries of status and power in favour of the dominant social group, and affect recipients' cognition and behaviour (see Eagly et al. 2000, Maass and Arcuri 1996, Menegatti and Rubini 2017). On the structural level, the norm according to which the prototypical human being is male is reproduced in many languages (Silveira 1980); feminine terms usually derive from the corresponding masculine form; and masculine nouns and pronouns are often used with a generic function to refer to both men and women (Menegatti and Rubini 2017). Here, we focus on the relation between gender, language and occupations; and also in this domain, a large body of work addresses stereotype-consistent language use (e.g. Heilman 2001, Gaucher et al. 2011), as well as investigating the influence of gender-fair language use (currently initiating heated debates in Slovenian professional and public spheres) in the context of job advertisements, or in societal perceptions of professions (Horvath and Sczesny 2016, Horvath et al. 2016).

Word embeddings

Word embeddings are vector representations of words: each word is assigned a vector of (typically) several hundred dimensions. These are usually obtained via training algorithms such as *word2vec* (Mikolov et al. 2013a) and *GloVe* (Pennington et al. 2014), which characterize the word based on the lexical context in which it appears. These representations improve performance in a wide range of automated text processing tasks, partly because they capture a degree of semantics: words that are similar or semantically related are closer together in vector space. They can also capture regularities beyond simple relatedness, such as analogies (Mikolov et al. 2013b); for example, the vector-space relation between *Madrid* and *Spain* is very similar to that between *Paris* and *France*.

This provides a way to analyse complicated concepts like gender. If we examine words which differ systematically in gender (e.g. *man*:*woman*; *son*:*daughter*), we expect the vector difference to be approximately the same (Pennington et al. 2014). We can discover gender correspondences via gender-based "analogies" (e.g. testing which word X is to *woman* as *king* is to *man*) by simple vector addition and subtraction (e.g. *king* – *man* + *woman* ≈ *queen*).

Word embeddings and biases

Being derived from naturally-occurring text, word embeddings also capture human biases, stereotypes and reflect social inequalities (Caliskan et al. 2017). Research on English word embeddings has shown examples of this effect: for example, the word *submissive* can be closer to *woman*, with *honourable* closer to *man* (Garg et al. 2017). This can be both because we often refer to men as being *honourable* directly, and because we refer to them in contexts in which we typically describe honourable things. Bolukbasi et al. (2016) showed that while this sometimes leads to rational outputs (e.g. in the analogy task *man*:*king* : : *woman*:*x;* the closest *x* corresponds to the vector of *queen*), it sometimes shows bias (e.g. *man*:*computer programmer* : : *woman*:*homemaker*). Caliskan et al. (2017) further demonstrated that embeddings contain biased associations (e.g. between math/arts and female/male terms), while Garg et al. (2017) used them to analyse gender stereotypes over time. Biases in word embeddings also influence

automated tools: Kiritchenko and Mohammad (2018) found that the majority of sentiment analysis systems tend to assign higher positivity to sentences involving some genders/races than others. Recently, efforts to decrease bias in embeddings have been made (e.g. Bolukbasi et al. 2016) - however, bias still persists to some extent (Gonen and Goldberg 2019). On the other hand, Nissim et al. (2019) warn that many studies may over-estimate bias.

Experiment with word embeddings in Slovene

Experimental setup

Inspired by the findings with English word embeddings described above, we also focus on occupations. In Slovene, gender for occupations is frequently expressed in morphology, e.g. *sociolog* (male) and *sociologinja* (female form) that we translate as *sociologist*$_{M}$ and *sociologist*$_{F}$, respectively.[2] Formulated as an analogy task, we look for gender analogies of occupations in both directions, finding the closest word embedding *x* for *woman*:*manager*$_{F}$: : *man*:*x* and vice versa for *man*:*manager*$_{M}$: : *woman*:*x*. The working hypothesis is that *x* should be the male or female version of the occupation, respectively, i.e. ženska:*menedžerka* : : *moški*:*menedžer* and *moški*:*menedžer* : : ženska:*menedžerka*. Slovene word embeddings were trained using word2vec on around 15 Gb of text (academic, news, books etc.).[3]

The female- and male-specific words for occupations, used in the experiment were taken from *the 1641st Regulation on the Introduction and Use of the Standard Classification of Occupations* (ULRS 28/1997), out of which we selected two groups of occupations where men and women had the highest quantitative hourly wage difference: (1) *Legislators, senior officials, managers* and (2) *Experts*, but also included occupations from the group with the smallest difference, i.e. *Officials* (Eurostat and SURS 2018, reporting data from 2014). Some occupations have only one version for both men and women (e.g. *vodja*) – these were treated as gender-neutral. Note that even if words for occupations have several synonyms (e.g. *dekanja*, *dekanica*, *dekanka*) – we used the one provided in the Regulation. From the initial 48 selected occupation pairs, for quantitative evaluation we removed the two gender-neutral pairs, as well as *corrector* (sl. *korektor, korektorica*) since the male form is a homograph for make-up corrector, resulting in 45 pairs. Two of the occupations (namely, *sekretar*/*sekretarka* and *tajnik*/*tajnica*) translate as *secretary* in English – we marked the higher-ranking occupation (sl. *sekretar* or *sekretarka*) as *secretary** and the lower ranking as *secretary*.

In experiments, the task was to find *x* in setting *man*:*occupation*$_{M}$: : *woman*:*x* (and vice versa), where *x* is the most similar word embedding (with the highest cosine similarity score). For each analogy, we included top 10 words or phrases.

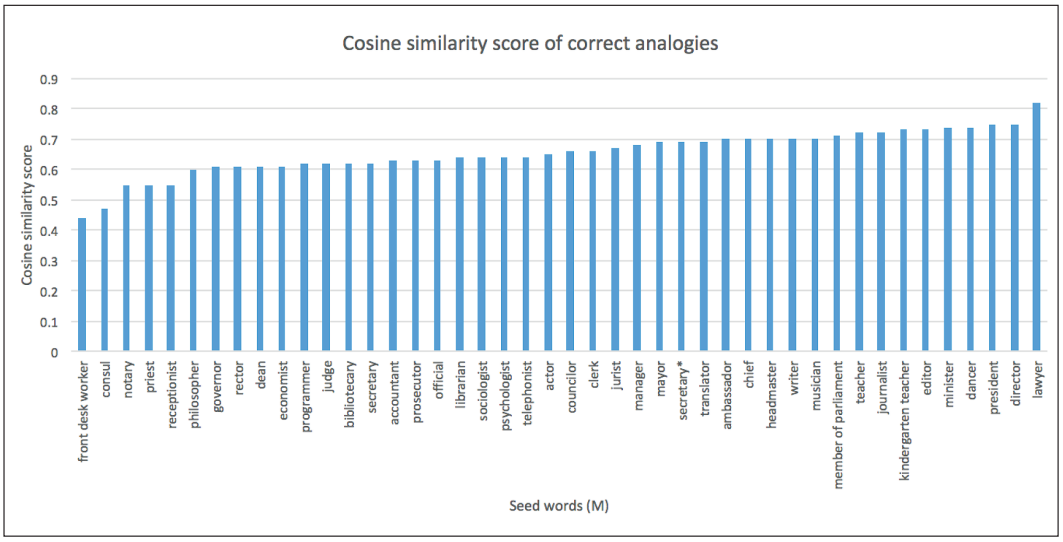
Experimental results and discussion

In general, the analogies followed the expected pattern. From 45 occupation word pairs, with female professions as seed words, male analogies were correct as the first hit in 71% and appeared in top 10 hits in 96% of cases. For the reverse task, the analogies were correct as the first hit in 87% of cases and appeared in top 10 hits in 98% of cases.
The correct match did not appear within the first 10 matches for two female word seeds—

2. In this paper, alternative word forms *(e.g. sociolog/inja or sociolog_inja)* are not taken into account.
3. The embeddings are the basis of kontekst.io (Plahuta 2019) and accessible upon request: https://kontekst.io/partnerstvo

receptionist$_F$ (sl. *recepcionistka*) and *front desk worker*$_F$ (sl. *informatorka*)—and once for male word seed (*attaché*). Examples when the match was not the first hit but was found in top 10 candidates include *secretary*$_F$ (sl. *tajnica*), where the first match for male equivalent was *boss*$_M$ (sl. šef), *priest*$_M$ (sl. *duhovnik*), where the first match was *nun* (sl. *nuna*), as well as *consul*$_M$, *notary*$_M$ (sl. *notar*) and *front desk worker*$_M$. The analogy *secretary*$_F$: *boss*$_M$ clearly stands out as an example, where the gender analogy expresses a hierarchical relation, and therefore reflects societal inequalities.

Figure 1. Cosine similarity score for correct female analogies for male occupation seed words[4].



For the correct matches in the analogy task, such as the pair *president*$_M$ (sl. *predsednik*) : *president*$_F$ (sl. *predsednica*), we computed the vector distances in similarity scores. For male specific occupations as seed words (Figure 1), the highest similarity score is observed for the occupations *lawyer* and *director*, while *front desk worker, consul, notary* and *priest* have the lowest score. It is interesting to observe that for two professions from the legal domain, *lawyer* is among the highest scored analogies, while *notary* is among the lowest; intuitively, this tells us that there are more differences in usage (and therefore perception) between *notary*$_M$ *and notary*$_F$ than there are between *lawyer*$_M$ and *lawyer*$_F$. In further work, it would be interesting to investigate in more detail where these differences lie and what they reflect; for this, (co-)occurrence corpus analysis of male and female forms and their contexts could be very informative. But even if the interpretation of these differences is not yet clear, it can serve as a starting point for investigating societal data. For example, according to the study *Mapping the Representation of Women and Men in Legal Professions Across the EU*, the distribution of notaries in Slovenia is imbalanced (cca. 40:60) in favour of women (Galligan et al. 2017,

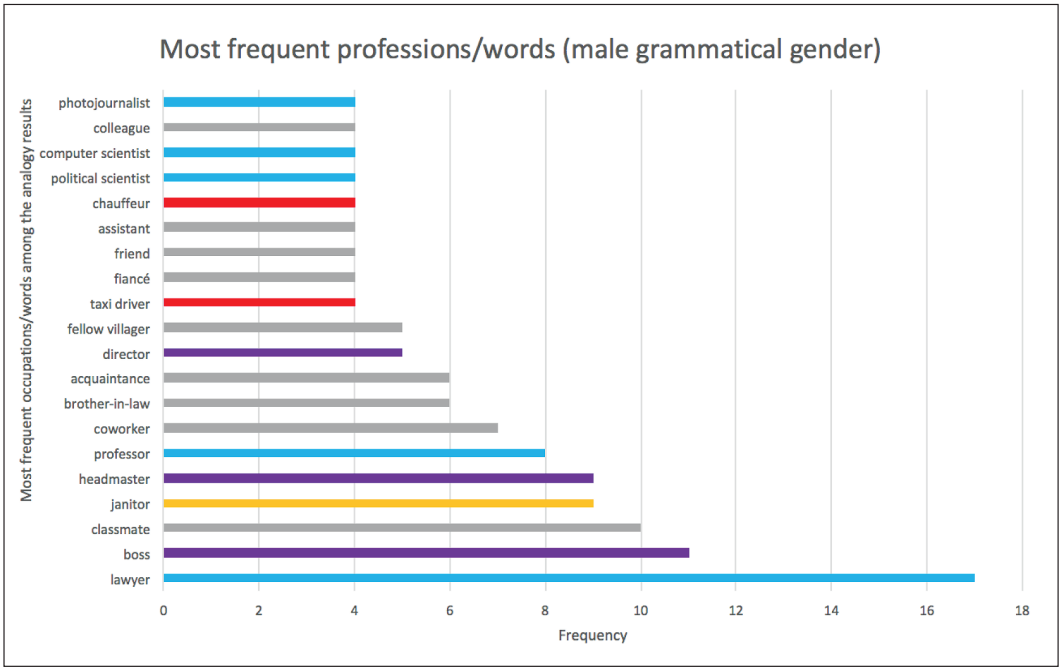
4. Occupation names in Slovene (as appearing in Figure 1): informator, konzul, notar, duhovnik, receptor, filozof, guverner, rektor, dekan, ekonomist, programer, sodnik, bibliotekar, tajnik, računovodja, tožilec, uradnik, knjižničar, sociolog, psiholog, telefonist, igralec, svetnik, referent, pravnik, menedžer, župan, sekretar, prevajalec, veleposlanik, načelnik, ravnatelj, pisatelj, glasbenik, poslanec, učitelj, novinar, vzgojitelj, urednik, minister, plesalec, predsednik, direktor, odvetnik.

69), as in the majority of former communist countries (a possible explanation being that the functions, prestige and income of a notary under communism was rather low and thus very different from the functions of a notary in a Western civil law country). On the other hand, the proportion of lawyers is imbalanced in favour of men (ibid., 64). However, distribution is certainly not the only factor, as for example, highly scored results also included occupations commonly associated with women (e.g. *kindergarten teacher* and *dancer*).

Not only first or correct matches, but also other analogues are interesting to analyse. For example, in analogues for *member of parliament* and *minister* more male proper names (politicians) occur. Also, for both directions, many words not related to the seed occupation were observed within the first 10 matches (e.g. *janitor*, *mechanic*, and *taxi driver* for males and *maid*, *housewife*, *servant*, *secretary*, *nurse*, *carer, cook* for females). Some of them correspond to popular occupations (see Vrabič Kek et al. 2016) that are mostly taken up by men (e.g. *mechanic*) or women (e.g. *nurse, secretary*). We therefore also analysed the top 20 male/female-specific words that appear within the first 10 matches of all analogies (see Figures 2 and 3). For males, there were many occupations that imply high social status (e.g. *lawyer*, two synonyms for *boss*, *director*, *headmaster*, *professor*, amounting to 50 counts altogether). Similar words appeared among the female-specific words (e.g. *lawyer*, *councillor*, two synonyms for *boss*, *vice-president*), but make up only 26 counts. The most common occupations (or words) among the male analogues were *lawyer* (sl. *odvetnik*) (17 examples), *boss* (sl. šef) (11), *classmate*-not an occupation (sl. *sošolec*) (10), *janitor* (sl. *hišnik*) (9), *headmaster* (sl. *ravnatelj*) (9). While *janitor* is nearly an exclusively male occupation, the other three are professions with high societal status, and belong to the categories with the highest wage difference per hour (above 2 eur). On the female side, the most common terms are *secretary* (sl. *tajnica*), *official* (sl. *uradnica*), *homemaker/housewife* (sl. *gospodinja*), *employee* (sl. *uslužbenka*) and *lawyer* (sl. *odvetnica*); here, with the exception of *lawyer,* all are occupations and roles with lower societal status and relatively small wage differences. The case of *housewife* is interesting, since it can mean both the occupation (*homemaker*; also found in the aforementioned regulation ULRS 28/1997) or can describe a stay-at-home woman. Given the presence of other words connected to house chores and care within the list (e.g. *maid*, *servant*$_F$, *hospital/care home worker*$_F$), even though none of our tasks in fact required analogies of these occupations, we can conclude that the connection between women and house chores was very much present in the original corpus on which the embeddings were trained.

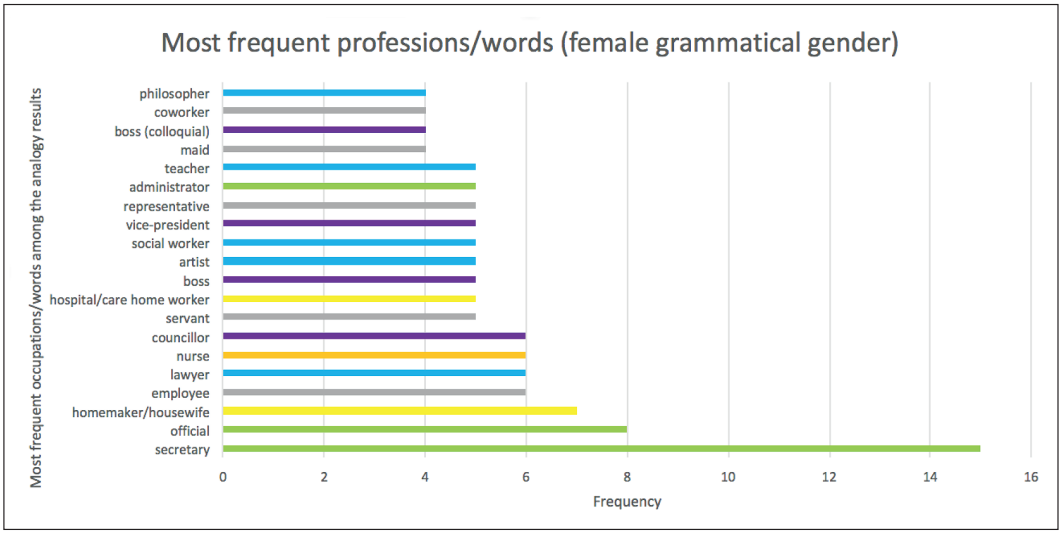
We also observed a few examples with stereotypical or even offensive analogies such as *stripper* (sl. *striptizeta*) for seed word *dancer*$_M$, or *gypsy* (sl. pej. *ciganka*) for *postman* (sl. *pismonoša*); the latter was not counted in quantitative results as it is a gender-neutral form.

Figure 2: Top 20 male specific words appearing within the first 10 matches of all analogies for female seed words[5].
Colour legend: (green – quantitative difference in wage per hour up to 0.49 eur; yellow – difference between 0.50 and 0.99 eur; orange – difference between 1.00 and 1.49 eur; red – difference between 1.50 and 1.99 eur; blue – difference between 2,00 and 2.49 eur; purple – difference over 2.50 eur) according to data from 2014 (Eurostat and SURS 2018). Words that represent non-specific professions *(e.g. assistant* (sl. pomočnik)) or not representing professions *(e.g. friend)* are marked with grey.



5. Occupation names in Slovene (as appearing in Figure 2): fotoreporter, stanovski kolega, računalničar, politolog, šofer, pomočnik, prijatelj, zaročenec, taksist, sovaščan, direktor, znanec, svak, sodelavec, profesor, ravnatelj, hišnik, sošolec, šef, odvetnik.

Figure 3: Top 20 female specific words appearing within the first 10 matches of all analogies for male seed words[6].
Colour legend refers to quantitative difference in wage per hour (see caption of Figure 2).



We have presented selected findings on gender bias in English word embeddings, and performed similar experiments on gender roles and occupations on Slovene.

By setting up a suitable analogy task – finding the female (or male) equivalent of a specified male (or female) profession – we show that a standard word embedding space for Slovene does exhibit gender regularities: in general, accuracy on the task is high. As expected, though, we also find that these regularities also capture stereotypes reflecting societal gender inequalities: the closest male analogue to $secretary_{F}$ (sl. *tajnica*) is found to be $boss_{M}$ (sl. šef); and the candidates for female analogue to $dancer_{M}$ (sl. *plesalec*) include *stripper* (sl. *striptizeta*). We also discovered that the most frequent close neighbours to the target occupation words seem to reflect similar stereotypes, with *nurse* closer to *woman* than to *man*, and with neighbours for male terms being more often high-status occupations, while those for female terms more often relate to low-status housework chores.

While these differences can be concretely measured, the interpretations thereof are currently rather more speculative; we expect this situation to improve with future developments of interpretability in NLP. However, we believe that these preliminary analyses clearly show the potential for embeddings-based analysis of gender as reflected in language and society.

Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper

6. Occupation names in Slovene (as appearing in Figure 3): filozofinja, sodelavka, šefica, služkinja, učiteljica, administratorka, predstavnica, podpredsednica, socialna delavka, umetnica, šefinja, strežnica, služabnica, svetnica, medicinska sestra, odvetnica, uslužbenka, gospodinja, uradnica, tajnica.

reflect only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

References

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James, Saligrama, Venkatesh, and Kalai, Adam (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of NIPS 2016.

Caliskan, Aylin, Bryson, Joanna J., and Narayanan, Arvind (2017): Semantics derived automatically from language corpora contain human-like biases. Science, 356 (6334): 183-186.

Eagly, Alice H., Wood, Wendy, and Diekman, Amanda B. (2000): Social role theory of sex differences and similarities: A current appraisal. In T. Eckes and H. M. Trautner (Eds.): The developmental social psychology of gender: 123-174. Mahwah: Lawrence Erlbaum.

Eurostat and SURS. (2018): 2.4 Plače. In Življenje moških in žensk v Evropi – statistični portret. Available at: https://stat.si/womenmen/bloc-2d.html (1. 6. 2019).

Galligan, Yvonne, Haupfleisch, Renate, Irvine, Lisa, Korolkova, Katja, Natter, Monika, Schultz, Ulrike, and Wheeler, Sally (2017): Mapping the Representation of Women and Men in Legal Professions Across the EU. Brussels: European Parliament.

Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, and Zou, James (2017): Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS, 115 (16).

Gaucher, Danielle, Friesen, Justin P., and Kay, Aaron C. (2011): Evidence that gendered wording in job advertisments exists and sustains gender inequality. Journal of Personality and Social Psychology, 101: 109-128.

Gonen, Hila, and Goldberg, Yoav (2019): Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. CoRR, abs/1903.03862.

Heilman, Madeline E. (2001): Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. Journal of Social Issues, 57 (4): 657-674.

Horvath, Lisa K., Merkel, Elisa F., Maass, Anne, and Sczesny, Sabine (2016): Does Gender-Fair Language Pay Off? The Social Perception of Professions from a Cross-Linguistic Perspective. Frontiers in Psychology, 6: 2018.

Horvath, Lisa K., and Sczesny, Sabine (2016): Reducing women's lack of fit with leadership? Effects of the wording of job advertisements. European Journal of Work and Organizational Psychology, 25: 316–328.

Kiritchenko, Svetlana, ad Mohammad, Saif M. (2018): Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. CoRR, abs/1805.04508.

Lakoff, Robin (1973): Language and Woman's Place. Language in Society, 2 (1): 45-80.

Maass, Anne, and Arcuri, Luciano (1996): Language and stereotyping. In C. N. Macrae, C. Strangor, and M. Hewstone (Eds.): Stereotypes and stereotyping: 193-226. New York: Guilford.

Menegatti, Michela, and Rubini, Monica (2017): Gender Bias and Sexism in Language. Oxford Research Encyclopedia of Communication. Available at: https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-470 (5. 9. 2019).

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey (2013a): Efficient Estimation of Word Representations in Vector Space. In Proceedings of ICLR 2013.

Mikolov, Tomas, Yih, Wen-tau, and Zweig, Geoffrey (2013b): Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL-HLT 2013.

Nissim, Malvina, van Noord, Rik, and van der Goot, Rob (2019): Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. CoRR, abs/1905.09866.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. (2014): GloVe: Global Vectors for Word Representation. In Proceedings of EMNLP.

Plahuta, Marko (2019). O slovarju. Available at: https://kontekst.io/o-slovarju (1. 6. 2019).

Silveira, Jeanette (1980): Generic masculine words and thinking. In C. Kramarae (Ed.): The voices and words of women and men: 165-178. Oxford: Pergamon.

ULRS 28/1997. Uradni list Republike Slovenije (št. 28/1997): 1641. Uredba o uvedbi in uporabi standardne klasifikacije poklicev. Available at: https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=199728&stevilka=1641 (5. 6. 2019).

Vrabič Kek, Brigita, Šter, Darja, and Žnidaršič, Tina (2016): Kako sva si različna: ženske in moški od otroštva do starosti. Ljubljana: SURS.

slovensko sociološko društvo

ZNANOST IN DRUŽBE
PRIHODNOSTI

SLOVENSKO SOCIOLOŠKO SREČANJE
Bled, 18. – 19. oktober 2019

Ljubljana, 2019

Izdajatelj:
Slovensko sociološko društvo
Kardeljeva ploščad 5, 1000 Ljubljana

Uredniki:
Miroljub Ignjatović, Aleksandra Kanjuo Mrčela, Roman Kuhar

Tehnični urednik:
Igor Jurekovič

Programski odbor:
Predsedstvo Slovenskega sociološkega društva

Recenzentke:
Anja Zalta, Alenka Švab in Veronika Tašner

Oblikovanje in prelom:
Polonca Mesec Kurdija

Korekture:
avtorji

Tisk:
Demat, d.o.o., Stegne 3, Ljubljana

Naklada:
150 izvodov

Prvi natis

Publikacija je dostopna tudi na elektronskem naslovu:
http://www.sociolosko-drustvo.si/.

Ljubljana, 2019

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

316(497.4)(082)

SLOVENSKO sociološko srečanje (2019 ; Bled)
Znanost in družbe prihodnosti / Slovensko sociološko srečanje, Bled 18.-19. oktober 2019 ; [uredniki Miroljub Ignjatović, Aleksandra Kanjuo Mrčela, Roman Kuhar]. - 1. natis. - Ljubljana : Slovensko sociološko društvo, 2019

ISBN 978-961-94302-3-1
1. Gl. stv. nasl. 2. Ignjatović, Miroljub
COBISS.SI-ID 302109696

Appendix D: Mitigating Gender Bias in Word Embeddings using Explicit Gender Free Corpus

Mitigating Gender Bias in Word Embeddings using Explicit Gender Free Corpus

David Hargrave
School of Electronic Engineering and Computer Science
Queen Mary University of London
d.r.hargrave@se19.qmul.ac.uk

***Abstract*—Words embeddings are the fundamental input to a wide and varied range of NLP applications. It has been shown that these embeddings reflect biases, such as gender and race, present in society and reflected in the text corpora from which they are generated, and that these biases propagate downstream to end use applications. Previous approaches to remove these biases have been shown to significantly reduce the direct bias, a measure of bias based on gender explicit words, but it was subsequently demonstrated that the structure of the embedding space largely retains indirect bias as evidenced by the spatial separation of words that should be gender neutral but are socially stereotyped on gender. This paper proposes a new method to debias word embeddings that replaces words in the training corpus that have explicit gender with gender neutral tokens, and creates the embeddings for these replaced words from the embedding of the gender neutral token post training utilising an added gender dimension. By design this method is able to fully mitigate direct bias and experiments demonstrate this. Experiments are also performed to investigate the effect on indirect bias, but generally are unable to achieve the reductions obtained by previous methods.**

I. Introduction

A word embedding is a compact representation of a word as a vector $\vec{w}$ in $\mathbb{R}^d$ with d usually between 50 and 300. An embedding space is thus a set of word embeddings. They are generated by algorithms such as GloVe (Pennington, Socher, and Manning 2014) and Word2Vec (Mikolov et al. 2013) that are trained on large text corpora.

Bolukbasi et al. (2016) identified that these embedding spaces contained gender bias. They defined a gender direction in the embedding space as the first principal component of the subspace spanned by a set of ten explicit gender word pairs[1]. The bias of a word embedding is then calculated as its projection, defined as the cosine similarity of normalized vectors, onto this gender direction. Using this metric they showed that gender biases defined by crowd workers were present in the embeddings, and conversely the crowd workers agreed with gender biases created from the embeddings.They coined the well known biased analogy found in the embedding space 'Man is to computer programmer as woman is to homemaker'

Caliskan, Bryson, and Narayanan (2017) demonstrated stereotyped biases, including gender bias, in word embeddings. They developed the Word-Embedding Association Test (WEAT) as analogous to the human Implicit Association Test (IAT). WEAT also used cosine similarity as a measure of correlation, and they were able to reproduce results from the IAT such as female names being more associated with family and arts as opposed to male names being more associated with career and mathematics.

They also showed that these biases propagate to downstream AI applications that use word embeddings. For example, in machine translation to English from a gender neutral language such as Turkish, "O bir doktor. O bir hemşire." translates to "He is a doctor. She is a nurse."

[1]Explicit gender words are those that explicitly define a gender such as he, woman, uncle, and queen. A gender pair is an equivalent pair of gender words such as he and she, man and woman etc.

II. Previous Work

As well as identifying the gender bias issue, Bolukbasi et al. (2016) also implemented two different algebraic methods to debias the word embeddings after training. Neutralize and Equalize adjusts the gender neutral word embedding vectors to be orthogonal to the gender direction and equidistant to both words in a gender pair (e.g. he and she). The less rigid Soften method seeks to maintain the structure of the embedding space by preserving pairwise inner products between all the word vectors whilst minimizing the projection of the gender neutral words onto the gender subspace.

Zhao et al. (2018) take the approach of modifying the cost function to debias the word embeddings during training, with the aim of forcing the gender component into the last dimension of the embedding vectors. They too identify a gender direction from a set of predefined gender pairs, as the average of the difference between the embeddings in each pair, excluding the last dimension. They modify the standard GloVe cost function to include additional terms, one to force the gender component for male and female words apart, and the second to make gender neutral words orthogonal to the gender direction.

Lu et al. (2018) proposed a method, Counterfactual Data Augmentation (CDA), to modify the text corpus before training. They identify a list of gender pair words, and duplicate the training corpus swapping words that occur in a gender pair with the other word in the gender pair, whilst retaining semantic correctness. The aim here is to create a gender balanced corpus on which to train the word embeddings.

Gonen and Goldberg (2019) devised a set of tests to demonstrate, that whilst Bolukbasi et al. (2016) and Zhao et al.

This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grant 825153 (EMBEDDIA).

(2018) did reduce the direct bias with respect to the definition as projection on to the gender direction, the embedding space still retained indirect bias, a structure between words that are not explicitly gendered but are socially stereotyped on gender, that can be used to infer gender based on the distance between vectors. Hall Maudslay et al. (2019) demonstrated a similar result for the Liu at al. (2018) method using a modification of the Gonen and Goldberg (2019) tests.

Hall Maudslay et al. (2019) also modified the CDA approach into Counterfactual Data Substitution (CDS), a technique that avoids duplication of text, by swapping gendered words pairs (from a list of 124 pairs) in situ with 50% probability. They also apply a technique called Names Intervention whereby names from the United States Social Security Administration (SSA) dataset are swapped in a manner that aims to preserve gender specificity[2] and frequency of use[3]. This approach is able to achieve a significant reduction in indirect basis using a modified method of the clustering technique defined by Gonen and Goldberg (2019). However this did not fully resolve the gender bias problem, and issues associated with it still remain.

In this paper an alternative pre-processing approach is proposed, whereby explicit gender is removed from the corpus before training. By design, this approach will yield equivalent results to the Neutralize and Equalize method of Bolukbasi et al. (2016) and will retain gender appropriate analogies. This paper aims to demonstrate these results and investigate the effect on indirect bias.

III. Methodology

Word embedding models such as GloVe (Pennington, Socher, and Manning (2014)) and Word2Vec (Mikolov et al. (2013)) use co-occurrence of words within a small window, to generate the word embedding vectors. Words such as he and she, king and queen, and man and woman, and given names have explicit gender, so that words that co-occur with these can inherit gender associations, and thus be biased, towards a particular gender. So an approach that can remove such associations from the training corpus may be able to mitigate this gender bias.

Word pairs such as he and she, king and queen, and man and woman essentially represent the same thing with the sole difference being gender. This suggests that the embedding vectors for these pairs of words should vary only in a gender direction. Given names also have gender (although to varying degrees) but are just labels assigned to people and really should not carry any other meaning, which suggests that they should cluster around some fixed vector and again vary only in a gender direction.

So the approach taken here is to substitute these gender pair words and names in the training corpus with gender neutral tokens before training. After the word embeddings have been trained, word embeddings for the individual gender pair words and names are created from the substituted token embedding vectors by the addition of an extra gender dimension. The values in this extra dimension are set to small non-zero values for the substituted words as described later, and to zero for all other words.

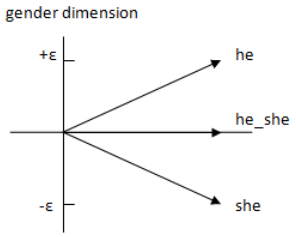


Fig. 1. Creation of word embedding for he and she from he_she. Here the he_she vector is shown as 1-d whereas in actuality it will be higher dimensional. The gender dimension is always 1-d.

For the gender pair words, a new token is created by combining the male word and the female word, separated by an underscore, e.g. man or woman are combined into man_woman. This new gender neutral token then replaces all occurrences of man or woman in the corpus. The list of gender pairs to be substituted is compiled as a subset of the definitional and equalize pairs used by Bolukbasi et al. (2016), and the list of male and female words used by Zhao et al. (2018).

Given names are replaced by a new token _name_ which is gender neutral. The identification of names uses the same source as used in the Names Intervention of Hall Maudslay et al. (2019), namely the United States Social Security Administration (SSA) dataset.

After the word embeddings have been created, embeddings for the substituted words are created from the embeddings of the substituting tokens. Firstly an extra dimension to represent gender is added to the embedding space.

For the substituted gender pair words, the male and female embedding vectors are created from the embedding of the combined token vector with the value in this extra dimension set to $+\epsilon$ for the male word in the pair and to $-\epsilon$ for the female word, for some small value of ϵ. An example is shown in Fig 1.

Embedding vectors for all substituted names are created as the embedding vector for the substituting token _name_ with a non-zero value in the added gender dimension. The value in the extra dimension is set to a value between $-\epsilon$ and $+\epsilon$, depending on the gender specificity of the name. The gender specificity is determined using the same method as Hall Maudslay et al. (2019). The SSA dataset, in addition to listing all names given in the US since 1880, also lists gender and frequency of use. It is thus possible to calculate a weighting between -1 (100% female) and +1 (100% male) and the value in the gender dimension is set to ϵ multiplied by this weighting. All other words will be considered gender neutral and have the gender dimension value set to 0.

[2]The degree to which a name is more male or female

[3]Names are swapped with other names that have similar frequency of use

This results in the difference between words in a gender pair being solely in the gender dimension e.g. $\vec{he}$-$\vec{she}$ is a vector of all zeros except for a value of 2ϵ in the gender dimension, whereas gender neutral words have a value of zero in the gender dimension. Thus this approach leads to the removal of direct bias;

- Gender neutral words are orthogonal to the gender pair directions such as $\vec{he}$-$\vec{she}$, which has been used as a definition of direct bias.
- Gender neutral words will be equidistant to both words in a gender pair e.g. man and woman.
- Correct gender associations for words in gender pairs, such as he is to she as man is to woman, will be present.
- And by controlling the size of ϵ it will be possible to ensure that the analogy test 'man is to surgeon as woman is to w?' will result in w=surgeon.

Experiments are performed to create the word embeddings using this prescribed methodology, and to demonstrate the removal of direct bias and investigate the effect on indirect bias.

IV. Experimental Setup

Wiki data is downloaded and preprocessed to create a training corpus. Original (biased) word embeddings are then created using the GloVe model. Gender pairs and names are then replaced in the training corpus with the substitution gender neutral token. The debiased embeddings are then created, after which the gender dimension is introduced. Word embeddings for the substituted gender pair words and names are then created from the embedding of the substitution token.

A. Data source

Wikipedia dumps are downloaded to create 3 separate corpora (500A, 500B and 500C) to train the word embeddings. Each corpus has approximately 500 million tokens. Details of the dumps used for these corpora are in Appendix A.

B. Data preparation

1) Preprocessing: Basic preprocessing is performed to prepare the corpus for training;

- split words separated by hyphen or forward slash
- remove words containing non-latin characters
- remove punctuation and special characters
- remove words that are not either all alphabetic or all numeric
- convert word to lowercase

Both sets of embeddings are created from the corpora that has had preprocessing applied.

2) Gender pairs: Gender pairs were collated from the definitional and equalize pairs of Bolukbasi et al. (2016) and the seed words of Zhao et al. (2018). Not all words were included.

- Pairs of words that were not considered to be exact matches that differ only in gender were removed e.g. fella and lady, beard and toque, and Catholic_priest and nun.
- Animal pairs were also removed because the selection is somewhat limited and arbitrary, and the language used to describe animals is considered to be less dependent on gender.
- Pairs where one of both words occur with low frequency were removed. However such pairs where the words could co-occur in a unique context were retained.

This resulted in a list of 77 unique gender pairs, which can be found in Appendix B.

3) Names: Names are retrieved from the same source as used by Hall Maudslay et al. (2019), namely the United States Social Security Administration (SSA) dataset. The dataset contains annual lists of all given names in the U.S. since 1880, including gender and count, e.g. the 2020 file has two entries for the name Taylor:

Taylor,F,1729
Taylor,M,456

From this, a gender specificity percentage (gsp) is calculated as:

$$gsp = (tmuc - tfuc)/(tmuc + tfuc) \quad (1)$$

where tmuc is the total male usage count and tfuc is the total female usage count.

4) Substitution: To prepare the corpus for the creation of the debiased word embedding vectors, individual male and female words in the corpus that occur in a gender pair, and names, are replaced by the appropriate substitution token.

The male and female words are replaced by the appropriate gender pair token e.g. 'he' or 'she' is replaced by 'he_she'.

Names are replaced with the '_name_' token. There are over 100,000 unique names in the SSA. Many of these names have a very low frequency usage count and thus name substitution was restricted to those that had been given at least 2,000 times in total, across both genders. This resulted in 7,082 names that were replaced. The usage counts in the SSA dataset show that these accounted for over 95% of given name usage. Seven names were removed from the list as they were also appeared as words in gender pairs.[4].

C. Word embedding creation

Word embeddings are created using the GloVe model. The parameters for the GloVe model are left unchanged from those in the demo.sh script downloaded from the Stanford GloVe website (Pennington (2014)), with the one exception being the vector size.

A set of original word embeddings are created from the original corpus with the vector size set to 200.

Substitution is then applied to the corpus, and a set of debiased word embeddings created, with a vector size of 199. The extra gender dimension is then added by increasing the vector size to 200, and setting the value in this gender dimension to zero.

[4]The removed names are Duke, Prince, Baron, Guy,King, Queen and Princess

Embedding vectors in the debiased space are then created for the substituted words, from the embedding vector of the substitution token. For words substituted by a gender pair, the individual word vectors are created from the gender pair vector with the value in the gender dimension set to $+\epsilon$ for the male word and to $-\epsilon$ for the female word. vectors for substituted names are created from the vector for the _name_ token. The value in the gender dimension is set to a value of ϵ multiplied by the gender specificity percentage calculated in equation 1.

The original and debiased embedding vectors are now ready to be used in the experiments.

V. Experimentation

In all experiments the gender dimension value ϵ is to 0.1. Both the original and debiased vectors are normalised for all experiments.

A. Direct bias

These 4 experiments are performed with the debiased embeddings and demonstrate the assertions that direct has been removed.

1) This test checks that all gender neutral words are unbiased. This is calculated as the sum of the absolute projection of all the gender neutral words (V_{GN}) onto the $\vec{he}$-$\vec{she}$ direction,
$$\sum_{w \in V_{GN}} \mid \vec{w}.(\vec{he} - \vec{she}) \mid$$
and is expected to be 0.
2) This test checks that gender neutral words are equidistant to both words in each gender pair. It calculates the difference between the distance to the male word and the distance to the female word, and sums this value for all gender neutral words (V_{GN}) and gender pairs (GP),
$$\sum_{w \in V_G N} \sum_{p \in GP} \|\vec{w} - \vec{p_m}\|_2 - \|\vec{w} - \vec{p_f}\|_2$$
where p_m is the male word in p and p_f is the female word in p. Again this is expected to be 0.
3) This test checks that appropriate analogies are present for the words in gender pairs. It uses the gensim (Řehůřek 2009) (version 3.8.3) 'most similar' function[5] to check that for each gender pair combination, p_1 and p_2, the question 'p_{1m} is to p_{1f} as p_{2m} is to w?' returns w=p_{2f} where p_{1m} and p_{2m} are the male words in the pairs and p_{1f} and p_{2f} are the female words in the pairs, e.g. for the pairs man_woman and king_queen, the question 'man is to king as woman is to w?' returns w=queen.
4) This test checks that the specific analogy 'man is to surgeon as woman is to w?' returns w=surgeon. Gensim is used to find the most similar word, and it's cosine similarity. Since gensim will not return any of the 3 input words, the cosine similarity of the these is also calculated. The word with the overall highest cosine similarity is deemed to be the best answer to the question. i.e. this test determines $\arg\max_{w \in V}(\vec{w}.(\vec{surgeon} - \vec{man} + \vec{woman}))$ where V is all words in the embedding space.

[5]Gensim uses cosine similarity to determine similarity, the word with the highest cosine similarity being the most similar. For normalised vectors this is equivalent to finding the nearest vector in Euclidean space.

B. Indirect bias

Gonen and Goldberg (2019) proposed 5 experiments to measure indirect bias, for which the code is supplied by the authors. They firstly reduce the embedding space to the 50,000 most common words, and from that remove the gender specific words used by Bolukbasi et al. (2016) and Zhao et al. (2018). The approach taken here is similar, but modified to remove the gender explicit words that have been substituted in the corpus prior to running the GloVe model, i.e. the gender pairs (which is a subset of their list), and the given names that have been substituted. This leaves only those words created by the GloVe model without intervention, and that are considered to be gender neutral, to be used in the experiments.

In all experiments the male/female gender bias of a word is defined as the projection of the word onto the $\vec{he}$-$\vec{she}$ direction in the original embedding space only (the same projection in the debiased space is now 0 for all gender neutral words).

1) Correlation between bias-by-projection and bias-by-neighbours: Implicitly gendered words, such as nurse or warrior, will no longer show direct bias in the debiased embedding space. This experiment suggests a measure of the indirect bias of a word as the correlation between the male/female bias of a word and the number of similarly biased words amongst it's nearest neighbours. Lower correlation will indicate less indirect bias.
2) Clustering: This experiment looks at how biased male and female words cluster together. It takes the 500 most biased male words and the 500 most biased female words, and performs k-means clustering (k=2), and then calculates the prediction accuracy of the clusters. The lower the accuracy[6], the more merged the male and female words have become, reducing indirect bias.
3) Professions: This experiment calculates the correlation between the gender bias of gender stereotypical professions, and the gender bias of the nearest 100 neighbours of the profession. The list of professions is taken from Bolukbasi et al. (2016). Less correlation indicates less indirect bias.
4) Classification: This experiment determines how well biased male and female words can be separated by an RBF-kernel SVM. It uses 5,000 words made up from the 2,500 most biased male words and the 2,500 most biases female word. It randomly takes 1,000 words to train an the classifier (500 from each gender) and then calculates the gender prediction accuracy on the remaining 4,000 words. Lower accuracy indicates less separation of the words and thus less indirect bias.

[6]This cannot be below 50% for 2 clusters

5) Association: This experiment replicates the gender related association experiments from Caliskan et al. (2017), but uses names as the gender identifier rather than gendered words (e.g. girl, her, brother). The experiments evaluate the association between male and female names, and 3 pairs of concepts that are considered to be gender biased, namely family and career, arts and maths, and arts and science. The experiments calculate the p-value. They do this by calculating the bias as the average absolute cosine similarity of the female names and female concepts, and male names and male concepts, and then calculating the same value for all combinations of names, and reports the percentage of times the combination of names has a higher bias than the original bias. The larger the p-value, the less likely there is an association between the names and concepts. Gonen and Goldberg (2019) use the terms 'art' and 'symphony' in experiments 2 and 3 as female concepts. These are also respectively a male and female name manipulated in the names processing, and so have been changed to 'theatre' and 'music' in experiment 2 and 'painting' and 'classics' in experiment 3.

   In addition Hall Maudslay et al. (2019) proposed adaptations to two of the Gonen and Goldberg (2019) experiments to measure indirect bias. The code for these experiments had to be modified to fit into the experimental framework provided by the Gonen and Goldberg (2019).
6) V-measure: This experiment reproduces the clustering experiment of Gonen and Goldberg (2019) with two variations.
   Firstly a different gender dimension is used to calculate bias in the original embedding space. It is defined as the first principal component of the subspace spanned by the difference between the word embeddings and the pair mean for each of the 23 pairs of words in the Google Analogy family test subset (GAF).
   $$\{p_m - \frac{p_m+p_f}{2}, p_f - \frac{p_m+p_f}{2} | p_m, p_f \epsilon p, p \epsilon GAF\}$$
   where p_m, p_f are the male and female words in the gender pair. And secondly tSNE (van der Maaten and Hinton (2008)) is done prior to the clustering.
7) Classification: This experiment reproduces the classification experiment of Gonen and Goldberg (2018) but uses the same definition of gender direction and bias as used in the V-measure experiment.

VI. Results

The set of experiments were run for all three datasets 500A, 500B and 500C. In addition experiments were run for two combinations of these datasets, 1000AB and 1000AC, both consisting of approx. 1 billion tokens (500A and 500B combined, and 500A and 500C combined respectively).

Results for the Gonen and Goldberg (2019) experiments are given along with those from their paper. Using their convention the results for Bolukbasi et al. (2016) are referred to as HARD-DEBIASED and for Zhao et al. (2018) as GN-GLOVE. Their results are based on word embeddings obtained from different datasets and thus the results are not directly comparable.

Results for the two Hall Maudslay et al. (2019) experiments are given along with the results for Counterfactual Data Substitution with Names Interventions (nCDS), and Counterfactual Data Augmentation (Lu et al. 2018) with Names Intervention (nCDA) from their paper as these were the two best performing techniques. Again these results are obtained from different datasets, albeit from Wikipedia.

A. Direct bias

These results apply to all datasets.

1) The projection of the gender neutral words onto $\vec{he} - \vec{she}$ is shown to be 0 as expected.
2) The gender neutral words are shown to be equidistant to the male and female words in the gender pairs as expected. The total difference is 0.
3) All combinations of gender pairs are shown to exhibit correct gender associations.
4) The test returns the word surgeon as the answer. As an example, gensim gives businesswoman[7] as the most similar word with a cosine similarity of 0.705, and the three input words have similarities of; woman 0.283, man 0.283, surgeon 1.000

Thus all four criteria are met demonstrating that direct bias has been removed.

B. Indirect bias

1) Correlation: The results show the Pearson correlation between the male bias of a word and the number of male biased words in it's 100 nearest neighbours.

	Before	After	Change
HARD-DEBIASED	0.741	0.686	-7.4%
GN-GLOVE	0.773	0.736	-4.8%
500A	0.729	0.680	-6.7%
500B	0.707	0.658	-6.9%
500C	0.714	0.664	-7.0%
1000AB	0.711	0.679	-4.5%
1000AC	0.703	0.672	-4.4%

   The results are consistent with those for HARD-DEBIASED and GN-GLOVE and show that this form of indirect bias still remains.

2) Clustering: The results show the cluster prediction accuracy.[8] 500A, 500C and 1000AC show a significant improvement over HARD-DEBIASED and GN-GLOVE, showing that there has been a siginifcant reduction of indirect bias. However, 1000AC does not show any improvement over the two smaller datasets, suggesting a limit may have been reached.

[7] Since gensim cannot return surgeon, this may be considered a more acceptable answer than nurse

[8] As there are two clusters, the accuracy cannot be below 50%

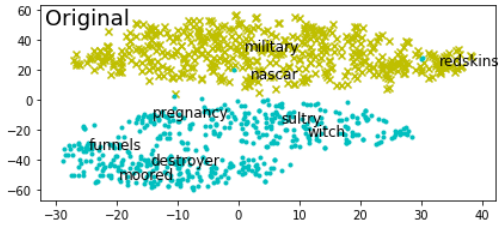


Fig. 2. Original clustering for the 500C dataset. Yellow represents the male words and cyan the female words.

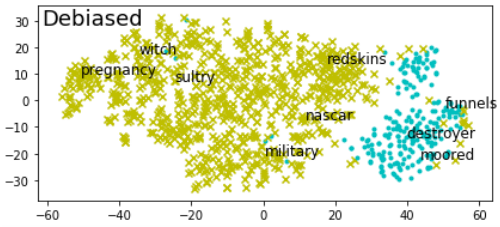


Fig. 3. Debiased clustering for the 500C dataset. The nautical words have formed a separate cluster (cyan) and the remaining female words have been incorporated into a single cluster with the male words.

	Before	After	Change
HARD-DEBIASED	0.999	0.925	-7.4%
GN-GLOVE	1.0	0.856	-14.4%
500A	1.0	0.695	-30.5%
500B	1.0	0.988	-1.2%
500C	1.0	0.727	-27.3%
1000AB	1.0	0.991	-0.9%
1000AC	1.0	0.705	-29.5%

500B and 1000AB perform poorly. Analysis shows that the most biased female words in 500A, 500C and 1000AC include a wide variety of nautical terms (e.g. destroyer, funnels, sank, torpedoed and drydock). This is presumably since ships are referred to with female pronouns. By substituting these female pronouns with gender neutral pairs, this association would be broken and it is observed that the nautical terms separate sufficiently from the other most biased female words to allow one of these groups to be incorporated into the male cluster. 500B and 1000AB do not exhibit this, so this could simply be an anomaly of the data. Fig 2 and Fig 3 shows this effect for 500C.

3) Professions: The results show the Pearson correlation between the male bias of a profession and the number of male biased words in it's 100 nearest neighbours.

	Before	After	Change
HARD-DEBIASED	0.747	0.606	-18.9%
GN-GLOVE	0.820	0.792	-3.4%
500A	0.817	0.788	-3.5%
500B	0.783	0.753	-3.8%
500C	0.794	0.745	-6.2%
1000AB	0.796	0.766	-3.8%
1000AC	0.766	0.720	-6.0%

500A and 500B perform similarly to GN-GLOVE, whereas 500C and 1000AC show some improvement but still well short of HARD-BIASED. This bias still remains.

4) Classification: The results show the prediction accuracy of the SVM classifier.

	Before	After	Change
HARD-DEBIASED	0.983	0.889	-9.6%
GN-GLOVE	0.987	0.965	-2.2%
500A	1.0	0.986	-1.4%
500B	1.0	0.979	-2.1%
500C	1.0	0.979	-2.1%
1000AB	1.0	0.990	-1.0%
1000AC	1.0	0.988	-1.2%

All datasets perform similarly to GN-GLOVE, but still short of HARD-BIASED. The bias still remains.

5) Association: The results show the p-values using the list of names supplied by Gonen and Goldberg (2019).

	Family-Career	Arts-Maths	Arts-Science
HARD-DEBIASED	0.0	0.0	0.047
GN-GLOVE	0.0	0.0	0.006
500A	0.524	0.476	0.476
500B	0.524	0.476	0.476
500C	0.524	0.476	0.476
1000AB	0.524	0.476	0.476
1000AC	0.524	0.476	0.476

The identical results are initally surprising. The word embeddings for the names have been created from the embedding of the _name_ token, based on gender specificity. This will result in the name embeddings having the same relative position in the gender dimension, and thus the projection of a word onto names will give the same relative (not absolute) values, in all experiments.
But the experiment is also dependent on the size of the differences between the projections. The method in which names are created means that when they are normalised the values in all other dimensions change very slightly, and so the differences in the projection of a word onto names are very small (<1e-06). This makes this experiment very sensitive to the names used.
For example changing one of the female names from Joan to Karen has little effect on GN-GLOVE (0.0, 0.0

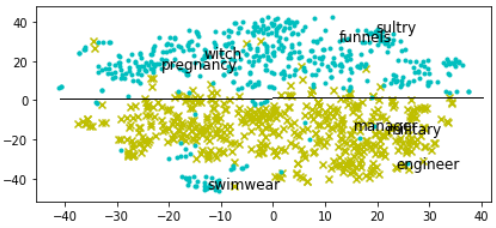


Fig. 4. tSNE mapping of the 500 most biased male and female words from which the V-measure is calculated for dataset 500B. The line is where the clusters will separate and thus shows where misclassification will occur

and 0.014)[9], but markedly (and consistently) changes the other results to (0.039, 0.960, 0.960). So rather than using a fixed set of names, a better approach is to use random samples of male and female names and average the results, and on dataset 500A for 500 iterations this gave results of 0.352, 0.671 and 0.671. Very similar results were seen for the other datasets. This suggests that this form of indirect bias has been removed.

6) V-measure:

	Before	After	Change
nCDA	1.0	0.594	-40.6%
nCDS	1.0	0.609	-39.1%
500A	0.963	0.262	-72.8%
500B	0.928	0.567	-38.9%
500C	0.940	0.354	-62.3%
1000AB	0.955	0.446	-53.3%
1000AC	0.903	0.683	-24.4%

There is an excellent reduction in cluster purity, with some datasets performing better than nCDA and nCDS[10]. In this experiment tSNE is performed before the kMeans clustering[11] However tSNE is highly sensitive to the order of input data such that if the order of the embeddings input to tSNE is switched, the V-measures change to: 500A-0.786, 500B-0.283, 500C-0.568, 1000AB-0.719, 1000AC-0.714. If tSNE is to be done first, it would be better to average over a large number of runs, but based on the results above the variance may be high.

Alternatively it may be preferable to perform kMeans without the prior tSNE, in which case the V-measures are: 500A-0.827, 500B-0.827, 500C-0.299, 1000AB-0.871, 1000AC-0.829.

7) Classification:

	Before	After	Change
nCDA	1.0	0.944	-5.6%
nCDS	1.0	0.889	-11.1%
500A	1.0	0.959	-4.1%
500B	1.0	0.957	-4.3%
500C	1.0	0.959	-4.1%
1000AB	1.0	0.963	-3.7%
1000AC	1.0	0.971	-2.9%

The results are slightly better than those from the previous classification experiment but not as good as nCDA and nCDS.

VII. Discussion

A. Results summary

This new approach to debiasing has, by design, removed direct bias, given a definition of the gender direction as $\vec{he} - \vec{she}$, Since, for all gender pairs, $\vec{p_1} - \vec{p_2}$ only has a value in the gender dimension, this result extends to all gender pairs, and any combination thereof.

It has also reduced clustering purity in both experiments. In the first experiment the results are mixed. Three datasets perform better than HARD-DEBIASED and GN-GLOVE, whereas the other two have hardly any effect. This could be related to the number of nautical terms present in the datasets and would be better assessed using a much larger and generalised dataset. With all five datasets it was observed that the cluster centers do move closer together, by up to 20%, showing that there is an degree of convergence of the most biased words.

The results of the second experiment are difficult to interpret given the sensitivity of tSNE. The results obtained without running tSNE first seem more reliable, and apart from the 500C figure, seem very consistent, although not performing as well as some of the datasets in the first experiment. In this experiment the definition of the gender direction is now more general than just $\vec{he} - \vec{she}$ and there are far fewer nautical terms in the most biased female words suggesting a more general degree of merging of the male and female words has occurred.

The method used to create name embeddings has also removed the connection between names and male/female concepts in the association experiments.

The results of the correlation experiment are similar to HARD-DEBIASED and GN-GLOVE. There is a small improvement over GN-GLOVE in the professions experiment, but well short of HARD-DEBIASED, and in the classification experiments the results are below that of HARD-DEBIASED, and nCDA and nCDS. It must be remembered that the results for HARD-DEBIASED, GN-GLOVE, nCDA and nCDS were produced on different datasets and so are not necessarily directly comparable.

So whilst this technique has performed reasonably well in removing indirect bias it generally has not been able to achieve superior performance to other techniques and much indirect bias still remains.

[9]It was not possible to rerun the HARD-DEBIASED test due to system constraints

[10]The results shown here are from a dataset taken from Wikipedia as this is considered more appropriate. nCDS did achieve a higher reduction of 58% on a dataset from Gigaword

[11]Hall Maudslay et al. (2019) state that 'For each biased embedding we then project these words into 2D space with tSNE (van der Maaten and Hinton (2008)), compute clusters with k-means, and calculate the clusters' V-measure (Rosenberg and Hirschberg, 2007).' The supplied code is consistent with this.

B. Methodology issues

All words that are not treated as gender pairs or names will be orthogonal to the gender direction and appear as gender neutral in the embedding space whilst clearly not all being so. This may have an effect on downstream applications.

Hall Maudslay et al. (2019) raise problems and limitations associated with an approach based on gender pairs. These include different spelling (mum v mom) and one-to-many associations (her v his and him, ladies v gentlemen and lords). In this paper, dad has been paired with mom, her with his, and ladies with lords since mom, his and lords are more frequent in the corpus but this approach is not wholly satisfactory.

There is also the issue of anatomical differences and related words. These can all imply gender (e.g. that person is pregnant) but do not pair off exactly (ovarian cancer is not the exact female equivalent of prostate cancer). Due to co-occurrence, these terms may well play a part in the gender separation of the embedding space.

C. Names polysemy

Many names are also words. Of the 7082 names substituted, 516 can be found in the US dictionary in the python enchant package (version 2.0.0). If substitution is extended to all 100,000+ names then 2,643 can be found in the dictionary. Substituting all 100,000+ names with the _name_ token reduces the quality of the embeddings as shown by the lower accuracy in the evaluation tests that follow the the creation of the embeddings by GloVe. Due to this, and since most names are used infrequently, only names with a total usage count count over 2,000 were selected, and it was seen that there was minimal impact on the evaluation accuracy. However this still leaves 516 words that will have an debiased embedding that is solely representative of the name, and other meanings will be lost, and this may reduce the performance of the embeddings in downstream applications. In the original embedding space these words will have an embedding that is a combination of the word and the name.

This limitation on the number of names that can be replaced without degrading the word embeddings means that there will be untreated names that can contribute to the gender separation of the embedding space, and additionally exhibit gender bias e.g. in the association experiment.

D. Experimental issues

The data preprocessing used here, together with the minimum vocabulary count in GloVe set to the default 5, created vocabularies with over 600K types (unique tokens) for datasets 500A and 500C, over 700K for 500B and over 1.0M for 1000AB and 1000AC. These vocabularies are too large. A separate experiment test run was done on dataset 500C with the minimum vocabulary count set to 10, which reduced the vocabulary size to 414K but performance was similar to the previous experiments except for the professions experiment which had a slightly larger reduction of -8.06%.

E. Observations

In the approach taken here, gender explicit terms are replaced with gender neutral terms, with the intention that this will give biased words a common reference in the cooccurrence matrix of Glove, i.e. whereas a female biased word would cooccur with 'she' and a male biased word with 'he', they would instead both cooccur with 'he_she', and the convergence of the clusters suggests that this is having the desired, although insufficient, effect. There are two points to make here. Firstly there are only 77 gender pairs and one name token that have been created to give this common cooccurrence, and a significant number of the gender pairs may not occur frequently in the training corpus. And secondly, the higher counts of cooccurrence of the gender pairs and especially the _name_ token will be penalised by the GloVe model as it both takes the log of the cooccurrence count and has a weighting function to limit the effect of frequent cooccurrences set at a value of 100 (Pennington, Socher, and Manning 2014). This may limit the effectiveness of this common cooccurrence to reduce indirect bias.

The approach taken here, and in the work of Bolukbasi et al. (2016), Zhao et al. (2018), Lu et al. (2018) and Hall Maudslay et al. (2019), is to focus on gender explicit terms and the removal of direct bias. However it may be that this approach is simply insufficient to address indirect bias.

It is interesting to note that the HARD-DEBIASED embeddings used by Gonen and Goldberg (2018), and which perform best in 3 of the experiments, are created using the Word2Vec model, whereas all other embeddings are created using GloVe. It may be worth creating HARD-DEBIASED embeddings from GloVe embeddings for a better comparison.

VIII. Further Work

The datasets used here are relatively small, so it is necessary to obtain results from a much larger dataset, and to produce results for the other methodologies on that dataset so that direct comparisons can be made. It would also be necessary to reduce the size of the vocabularies used, either through improved preprocessing or use of the minimum vocabulary count setting in GloVe.

Evaluation of the debiased embeddings in gender sensitive downstream applications should be investigated. This could include applications such as sentiment analysis on named individuals, coreference resolution, or CV and application form processing.

Modification of the GloVe cost function to allow for a greater contribution from the gender pair and _name_ tokens may be justified in that these terms represent more than a single word/name, and could lead to reduced indirect bias.

A solution to the issue of name polysemy needs to be found.

IX. Acknowledgements

I would like to thank Rowan Hall Maudslay for kindly providing the code for the experiments in Hall Maudslay et al. (2019).

REFERENCES

Bolukbasi, Tolga et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186. ISSN: 0036-8075. DOI: 10.1126/science.aal4230. eprint: https://science.sciencemag.org/content/356/6334/183.full.pdf. URL: https://science.sciencemag.org/content/356/6334/183.

Gonen, Hila and Yoav Goldberg (June 2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614. DOI: 10.18653/v1/N19-1061. URL: https://aclanthology.org/N19-1061.

Hall Maudslay, Rowan et al. (Nov. 2019). "It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5267–5275. DOI: 10.18653/v1/D19-1530. URL: https://aclanthology.org/D19-1530.

Lu, Kaiji et al. (2018). "Gender Bias in Neural Natural Language Processing". In: *CoRR* abs/1807.11714. arXiv: 1807.11714. URL: http://arxiv.org/abs/1807.11714.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., pp. 3111–3119.

Pennington, Jeffrey (2014). *GloVe: Global Vectors for Word Representation*. URL: https://nlp.stanford.edu/projects/glove/ (visited on 04/30/2021).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Řehůřek, Radim (2009). *Topic Modelling for Humans*. URL: https://radimrehurek.com/gensim/ (visited on 03/15/2021).

van der Maaten, L.J.P. and G.E. Hinton (2008). "Visualizing High-Dimensional Data Using t-SNE". English. In: *Journal of Machine Learning Research* 9.nov. Pagination: 27, pp. 2579–2605. ISSN: 1532-4435.

Zhao, Jieyu et al. (Oct. 2018). "Learning Gender-Neutral Word Embeddings". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4847–4853. DOI: 10.18653/v1/D18-1521. URL: https://aclanthology.org/D18-1521.

Appendix A: Data Sets

Dataset 500A:

Downloaded from https://dumps.wikimedia.org/enwiki/20210601/ [12]

enwiki-20210601-pages-articles1.xml-p1p41242.bz2
enwiki-20210601-pages-articles10.xml-p4045403p5399366.bz2
enwiki-20210601-pages-articles12.xml-p7054860p8554859.bz2
enwiki-20210601-pages-articles13.xml-p10672789p11659682.bz2
enwiki-20210601-pages-articles14.xml-p11659683p13159682.bz2
enwiki-20210601-pages-articles15.xml-p14324603p15824602.bz2
enwiki-20210601-pages-articles16.xml-p17460153p18960152.bz2
enwiki-20210601-pages-articles17.xml-p20570393p22070392.bz2
enwiki-20210601-pages-articles18.xml-p23716198p25216197.bz2
enwiki-20210601-pages-articles19.xml-p27121851p28621850.bz2

Dataset 500B:

Downloaded from https://dumps.wikimedia.org/enwiki/20210701/ and https://dumps.wikimedia.org/enwiki/20210720/ [12]

enwiki-20210701-pages-articles20.xml-p32808443p34308442.bz2
enwiki-20210701-pages-articles20.xml-p34308443p35522432.bz2
enwiki-20210701-pages-articles21.xml-p35522433p37022432.bz2
enwiki-20210701-pages-articles21.xml-p37022433p38522432.bz2
enwiki-20210701-pages-articles21.xml-p38522433p39996245.bz2
enwiki-20210701-pages-articles22.xml-p39996246p41496245.bz2
enwiki-20210701-pages-articles22.xml-p41496246p42996245.bz2
enwiki-20210701-pages-articles22.xml-p42996246p44496245.bz2
enwiki-20210701-pages-articles22.xml-p44496246p44788941.bz2
enwiki-20210701-pages-articles23.xml-p44788942p46288941.bz2
enwiki-20210701-pages-articles23.xml-p46288942p47788941.bz2
enwiki-20210720-pages-articles23.xml-p47788942p49288941.bz2
enwiki-20210720-pages-articles23.xml-p49288942p50564553.bz2
enwiki-20210720-pages-articles24.xml-p50564554p52064553.bz2
enwiki-20210720-pages-articles24.xml-p52064554p53564553.bz2
enwiki-20210720-pages-articles24.xml-p53564554p55064553.bz2
enwiki-20210720-pages-articles24.xml-p55064554p56564553.bz2

Dataset 500C:

Downloaded from https://dumps.wikimedia.org/enwiki/20210701/ and https://dumps.wikimedia.org/enwiki/20210720/ [12]

enwiki-20210701-pages-articles27.xml-p66975910p68108549.bz2
enwiki-20210720-pages-articles26.xml-p62585851p63975909.bz2
enwiki-20210720-pages-articles27.xml-p63975910p65475909.bz2
enwiki-20210720-pages-articles27.xml-p65475910p66975909.bz2
enwiki-20210720-pages-articles27.xml-p66975910p68286200.bz2
enwiki-20210720-pages-articles3.xml-p151574p311329.bz2
enwiki-20210720-pages-articles4.xml-p311330p558391.bz2
enwiki-20210720-pages-articles5.xml-p558392p958045.bz2
enwiki-20210720-pages-articles6.xml-p958046p1483661.bz2
enwiki-20210720-pages-articles7.xml-p1483662p2134111.bz2

[12] All links correct as of 31st July 2021

Appendix B: Gender pairs

From Bolukbasi et al. (2016b) definitional pairs

man	woman
boy	girl
he	she
father	mother
son	daughter
guy	gal
male	female
his	her
himself	herself

From Bolukbasi et al. (2016b) equalisation pairs

monastery	convent
spokesman	spokeswoman
monk	nun
dad	mom
men	women
councilman	councilwoman
grandpa	grandma
grandsons	granddaughters
uncle	aunt
husbands	wives
husband	wife
boys	girls
brother	sister
brothers	sisters
businessman	businesswoman
chairman	chairwoman
congressman	congresswoman
dads	mums
boyfriend	girlfriend
fatherhood	motherhood
fathers	mothers
fraternity	sorority
lord	lady
lords	ladies
grandfather	grandmother
grandson	granddaughter
king	queen
males	females
nephew	niece
prince	princess
schoolboy	schoolgirl
sons	daughters

Additional from Zhao et al. (2018)

countryman	countrywoman
actor	actress
bachelor	bachelorette
papa	mama
governor	governess
sir	madam
househusband	housewife
god	godess
groom	bride
emperor	emperess
landlord	landlady
duke	duchess
fiance	fiancee
stepfather	stepmother
policeman	policewoman
paternity	maternity
masseur	masseuse
mr	mrs
headmaster	headmistress
czar	czarina
stepson	stepdaughter
homosexual	lesbian
waiter	waitress
heir	heiress
monks	nuns
hero	heroine
abbot	abbess
widower	widow
baron	baroness
host	hostess
godfather	godmother
priest	priestess
patriarch	matriarch
actors	actresses
paternal	maternal
kings	queens

Appendix E: Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 39–49
DOI: 10.17645/mac.v8i3.3022

Article

Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting

Leo Leppänen [1,*], Hanna Tuulonen [2] and Stefanie Sirén-Heikel [3]

[1] Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland; E-Mail: leo.leppanen@helsinki.fi
[2] Swedish School of Social Science, University of Helsinki, 00014 Helsinki, Finland; E-Mail: hanna.tuulonen@helsinki.fi
[3] Media and Communication Studies, University of Helsinki, 00014 Helsinki, Finland; E-Mail: stefanie.siren-heikel@helsinki.fi

* Corresponding author

Submitted: 15 March 2020 | Accepted: 10 June 2020 | Published: 10 July 2020

Abstract
In this article we consider automated journalism from the perspective of bias in news text. We describe how systems for automated journalism could be biased in terms of both the information content and the lexical choices in the text, and what mechanisms allow human biases to affect automated journalism even if the data the system operates on is considered neutral. Hence, we sketch out three distinct scenarios differentiated by the technical transparency of the systems and the level of cooperation of the system operator, affecting the choice of methods for investigating bias. We identify methods for diagnostics in each of the scenarios and note that one of the scenarios is largely identical to investigating bias in non-automatically produced texts. As a solution to this last scenario, we suggest the construction of a simple news generation system, which could enable a type of analysis-by-proxy. Instead of analyzing the system, to which the access is limited, one would generate an approximation of the system which can be accessed and analyzed freely. If successful, this method could also be applied to analysis of human-written texts. This would make automated journalism not only a target of bias diagnostics, but also a diagnostic device for identifying bias in human-written news.

Keywords
algorithmic journalism; automated journalism; bias; diagnosis; journalism; news automation

Issue
This article is part of the issue "Algorithms and Journalism: Exploring (Re)Configurations" edited by Rodrigo Zamith (University of Massachusetts–Amherst, USA) and Mario Haim (University of Leipzig, Germany).

1. Introduction

In the current news media landscape, examining and acknowledging underlying bias is an important step in strengthening newswork and rectifying trust in journalism. As media is becoming reliant on metrics and personalization, striving for balance in issues such as gender, race, age, socioeconomic status, and story topics become increasingly poignant. Particularly when considering the expectations of the public of news as a representation of 'reality' (Reese & Shoemaker, 2016, p. 393). While working towards this goal, it is somewhat common to view automated journalism as a savior: an 'unbiased,' 'fair' and 'objective' decision-making system in comparison to the seemingly biased decision-making of humans. From this point of view, increased automation in the newsroom sounds like a match made in heaven, as newsrooms strive to be bastions of objectivity (Mindich, 2000, p. 1). As such, it comes as no surprise that many newsrooms are either already employing automated journalism or are interested in doing so (Sirén-Heikel, Leppänen, Lindén, & Bäck, 2019).

While the literature on automated journalism has presented various partially conflicting definitions (cf. Graefe, 2016), a very useful one is provided by Dörr (2016) and Caswell and Dörr (2018), who approach au-

tomated journalism through the technology employed. In their view automated journalism is about the employment of Natural Language Generation methods for producing news text. Natural language generation is a "subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable text in English or other human languages from some underlying nonlinguistic representation of information" (Reiter & Dale, 1997, p. 57). As such, Caswell and Dörr's (2018) definition explicitly excludes, for example, systems that produce summaries of news content written by other humans.

In this article, we use the term automated journalism along the lines of Caswell and Dörr (2018). In our view, automated journalism is the act of automatically producing a complete or near-complete news text from some underlying data. We include the qualifier 'near-complete' as a conscious acknowledgement of the view that a human can—and perhaps should—be included in the journalistic process of publishing. In practice, this means that our definition includes systems that produce story 'blanks,' raw textual material which already contain the main beats of the story but need further human editing before they are ready for audiences.

Irrespective of the precise definition of automated journalism, we believe it to be important to inspect the technology critically. As pointedly demonstrated by the now (in)famous analysis of automated prediction of recidivism (Angwin, Larson, Mattu, & Kirchner, 2016), algorithmic biases can have substantial effects. If the algorithms are viewed with an assumption of fairness, they present a danger of entrenching and hiding pre-existing biases. In the context of journalism, a profession and product defined largely by ideals such as objectivity, neutrality and factuality, it is crucial that unwanted biases are not allowed to entrench themselves unnoticed in the language and the content of news.

Other authors have previously researched both how algorithms can be investigated for journalistic purposes (Diakopoulos, 2015), described how algorithms involved in newswork could be made transparent (Diakopoulos & Koliska, 2017) and provided descriptions of how automation can help reduce bias in reporting (Fischer-Hwang, Grosz, Hu, Karthik, & Yang, 2020). Similarly, some technical works have investigated methods for identifying bias in non-journalistic contexts (e.g., Caliskan, Bryson, & Narayanan, 2017; Knoche, Popović, Lemmerich, & Strohmaier, 2019). In this article, we synthesize how these methods and ideas apply to diagnosing automated journalism itself for bias.

Such diagnoses can serve multiple purposes. First, they would quite naturally be of interest to researchers, as they would increase our understanding of the news media. Second, they would be of interest to third-party interest groups as a method for highlighting potential biases against any one of multiple demographics. Third, they present an opportunity to the newsrooms themselves to highlight the results of the audits as benchmarks or as societal commentary. Statistics on the gender distribution in news stories is already used by some news organizations for benchmarking (Helsingin Sanomat, 2018).

In relation to bias in news journalism, bias has conventionally been studied from the perspective of an autonym to objectivity, having adverse effects on the journalistic ethos to report reality truthfully, and as a symptom of partisanship (Hackett, 1984). As journalism is conceptualized as the fourth estate in democratic societies, bias has largely been tied to politics and ideology, editorial policy and individual journalists. The complexity of journalistic bias has gained a new dimension with digitalization. The shift towards mobile and the changes in audience behavior has increased the role of the audience, affecting news values and journalistic work (Harcup & O'Neill, 2016; Kunert & Thurman, 2019). Personalization, in effect a form of bias, has become a strategy for media organizations and platforms for creating customer value. Catering for audience tastes based on implicit or explicit user information can also increase the value for automated news, for example based on location, as suggested by Plattner and Orel (2019). However, as Kunert and Thurman (2019) found in their longitudinal study, most news organizations remain committed to exposing their audience to a diversity in news stories, reaffirming the prevailing framing of quality journalism.

Distinguishing between 'acceptable' bias, such as exhibited in personalized sports news, and 'unacceptable' bias, e.g., favoring certain ethnicities, is a value ridden process. Both are examples of 'selectivity,' as suggested by Hofstetter and Buss (1978, p. 518), or more generally framing (see Entman, 1993; Scheufele, 1999). Only shared values decide that one is acceptable and the other is not. Encoding such values exhaustively into any automated procedure is extremely difficult. It is unlikely that automated methods will be able to make this distinction outside of the most blatant cases. As such, when we refer to 'detecting bias,' 'causing bias,' etc., we are in fact talking about biases of 'undetermined polarity,' meaning that additional human analysis is required to determine whether the potential biases detected are acceptable or not. Nonetheless, due to the effects of media on audience perceptions, consciousness of bias and embedded values in automated journalism is of paramount importance.

2. Bias in Automated Journalism

Despite increased media attention, the term 'algorithmic (un)fairness' is still unfamiliar for many (Woodruff, Fox, Rousso-Schindler, & Warshaw, 2018, pp. 5–7). This is understandable as the 'unbiasedness' and 'fairness' of algorithms is often expressed as a selling point of automation: The prospect of a perfectly fair and objective computer replacing the biased human as the maker of hiring decisions, arbitrator of loan applications, and judge of those accused of crimes is very enticing.

Automated journalism has mostly been employed in settings where the objectivity standard can be considered the highest, such as weather reports (Goldberg, Driedger, & Kittredge, 1994) and financial news coverage (Yu, 2014). While automation has since been applied to domains where news media often produce more subjective commentary, such as elections and sports (Diakopoulos, 2019, p. 107), to the best of our knowledge even in these domains the systems tend to be applied to what we consider the objective side of the topic, reporting results rather than analysis.

While this positioning of automated journalism in the larger journalistic field is clearly driven by technology to some degree (i.e., the technology being unsuitable for other, more subjective, story types; see e.g., Stray, 2019), it seems that the view that objectivity is the best aspect of automation is also an influencing actor. The views of the media seem to be exemplified by the words of an editor of a regional media company, who stated that automatically produced stories represent "facts…and figures, not someone's manipulated interpretation" (Sirén-Heikel et al., 2019, p. 56). To us, such beliefs indicate two crucial assumptions: that removing the individual—or the first level of hierarchy of influences (Reese & Shoemaker, 2016)—is sufficient to remove bias, and that using automation indeed removes the effect of the individual. We will return to these assumptions in the conclusions of this article.

As increasingly acknowledged both within and without computer science, the use of algorithms is not a panacea to removing bias from society, if such a thing is feasible at all. In fact, automated systems are increasingly recognized as reflecting existing societal biases (Selbst, boyd, Friedler, Venkatasubramanian, & Vertesi, 2019) and due to the 'objective' imagery associated with them they might further systematize these biases. At the same time, it is hard to define what, exactly, it would mean for an algorithm to be unbiased or 'fair' (Woodruff et al., 2018, p. 1), with some notions of algorithmic fairness even being fundamentally incompatible with each other (Friedler, Scheidegger, & Venkatasubramanian, 2016, p. 14). As an example of the complexities of the topic, consider whether a system that simply reflects some underlying societal bias—and would automatically stop doing so if the societal bias was removed—is by itself biased? Due to these difficulties in defining what, precisely, is fair and unbiased, we do not focus our efforts on identifying what is unbiased or proscribing how the world should be. Instead we will next consider a few examples of cases where a system for automated journalism is either clearly biased, or at least raises the question of whether the system or the society it is employed in is biased.

We base our analysis on the observation that, in very broad conceptual terms, natural language generation can be thought of as consisting of three major subprocesses: deciding what to say, deciding how to say it, and actually saying it (Gatt & Krahmer, 2018, p. 84; Reiter & Dale, 2000, p. 59). The distinction between the last two steps is that whereas the second step decides e.g., what words to use, and in which grammatical forms, the actual inflection is done at the third step. It seems clear to us that if a system for automated journalism results in biased output when starting from data considered objective, the bias must have been introduced in either the first or the second step.

At the same time, whether based on human-written rules or machine learning, a system for automated journalism can also produce biased output text if the system inputs are biased. For example, an ice hockey reporting system will only produce news about the male leagues if it is never provided the results for the female leagues. Bias resulting from biased input is, however, distinctly different from biases built into the automated systems, with the operative difference being which part of the process must be modified to address the bias. Any system will misfunction when presented with incorrect inputs, or as the saying goes: 'garbage in, garbage out.' While a system receiving incorrect information indubitably reflects badly on the journalists and editors responsible for the system, it does not necessarily indicate that the system itself is malfunctioning. For this reason, in order to understand the weaknesses of the system, we must first focus on whether it malfunctions in the case of correct, i.e., unbiased, inputs. As such, going forward with our analysis, we will assume that the system is receiving correct, unbiased inputs.

As noted above, biases introduced by the system must be related to either content selection or the language used in the text. We will now consider the kinds of biases that could be introduced in both steps separately.

2.1. Bias in News Content Selection

With bias in news content selection, we refer to any phenomenon where the inclusion and exclusion of pieces of information from a news article reflects a potential bias. A real-life example of such a bias is described by Hooghe, Jacobs, and Claes (2015), who observe that female members of parliament received less speaking time than their male colleagues in Belgian media. Similar phenomena have been observed, for example, in sports reporting, where the coverage of male sports significantly eclipse the coverage of female sports (Eastman & Billings, 2000) and in reporting about same-sex marriages, where male sources were more likely to be quoted than female (Schwartz, 2011).

Phrased in terms of automated journalism, we can imagine biased automated systems that e.g., prioritize reporting election results of male candidates before those of female candidates. However, it is important to note that simply quoting more male politicians than female politicians does not necessitate that the automated system has a gender bias. Instead, it might be simply reflecting underlying societal factors and biases: If there are 99 male politicians to one female politician, a system ran-

domly picking a candidate to quote would mostly quote males. A more nuanced analysis is needed in such cases.

These content selection biases can, however, be more subtle and less obvious. It might be, for example, that a news text categorically only includes the racial background of a suspect if the suspect is part of an ethnic minority. Or similarly, reporting of a car crash might only mention the gender of the driver if they are female. In both cases, such reporting could entrench prior reader biases, either affirming their biased beliefs (those who are part of an ethnic minority commit more crimes, women are worse drivers) or not presenting contradicting evidence (a suspect of unspecified ethnicity committed a crime, a driver of unspecified gender crashed).

These examples show that bias can result not from just exclusion of information (i.e., protected classes being ignored or underrepresented in reporting), but also from highlighting the membership in a protected class.

2.2. Bias in News Language

It is also possible for the language of the news to be biased even in cases where the information content itself is not necessarily so. For example, Eastman and Billings (2000, p. 208) observe a tonal difference in sports reports, where male athletes were discussed in an enthusiastic tone, while female athletes were discussed in a derogatory tone.

Such linguistic bias can manifest in the minor difference in the nuance of the words that are used in the news text. For example, there is significant tonal difference in whether a car accident is described using language where the actor of the event is the pedestrian ('a pedestrian ended up being hit by a car'), the car ('a car ran over a pedestrian') or the driver of the car ('a driver ran over a pedestrian'). Minor changes in the lexical choice presents the driver of the car as having a passive role in the event, almost making them an observer, even if the facts of the event place most of the blame on the driver. Seemingly minor choices such as these can be seen as biased against those of lower socioeconomical status, who are less likely to own a car and more likely to be pedestrians.

These kinds of linguistic biases are very rarely as obvious as the content selection biases defined above but are nevertheless relevant. Minor changes in lexical choice can have significant effect. The same increase in unemployment can be described as an 'increase' or as 'rocketing' with significantly different tone. Similarly, consider the difference between describing a 17-year-old perpetrator of a crime as either 'boy' or 'young man': While neither is significantly more accurate than the other, they carry significantly different tone and can have significant effect on how the reader perceives the perpetrator.

There is nothing inherent to automated journalism that would prevent such biased language from being produced by an automated system, just like there is nothing inherent to the automation that would prevent systems from having biases in content selection. Next, we consider the mechanisms that would allow such biases to appear in the text produced by automated journalism.

3. The Mechanisms for Biased Automated Journalism

The previous sections highlighted ways in which the output of a system for automated journalism could be biased. It did not, however, address the mechanism by which such biases end up in the system. We now turn to this question.

Automated journalism, as in the automated production of news texts, can fundamentally be achieved by two technical methods (Diakopoulos, 2019). The first of these is via algorithms consisting of human-written rules that directly govern the actions of the system. The second is via algorithms that learn the rules from examples provided by the system creators, also known as machine learning. We will next discuss both approach in turn, with special focus on how biases might end up being encoded in such systems.

3.1. Bias in Rule-Based Systems

Rule-based systems for automated journalism are based on explicit rules programmed by human programmers, such as 'start an article on election results by mentioning who is now the largest party, unless some party lost more than 25% of their seats, in which case discuss them first.' Such rules, however, can be implemented using various technical methods and are best defined by the common factor that they are not automatically learned from examples. As these systems are, fundamentally, driven by rules and heuristics produced directly by humans, the principal reason for these systems to produce biased content is by the human-produced rules being biased.

Commercial actors providing systems for content creation or distribution, particularly those involving automation or machine learning, tend to keep their systems' details largely hidden from the research community. Naturally, this also holds true for systems used for automated journalism. However, interviews with media industry representatives indicate that most of the systems employed in the real-world newsrooms are indeed rule-based, rather than based on complex machine learning (Sirén-Heikel et al., 2019). Based on the limited evidence available, such as the few open source systems (e.g., Yleisradio, 2018), these systems are often based on what can be described as 'story templates.' These templates are, in broad terms, the algorithmic equivalent of the combination of a Choose Your Own Adventure book and a Mad Libs word game. The software inspects the input data, and based on human-written rules, selects which spans of text to include in the story and in which order. These 'skeleton' text spans contain empty slots, where values from the input are then embedded to produce the textual output of the story. While significantly more complex rule-based methods exist, espe-

cially in academia (see, e.g., Gatt & Krahmer, 2018, for an overview), the decree to which they have entered use in the industry is not clear to us.

Irrespective of the technical details of the system, the important factor in these types of rule-based systems is that on a fundamental level they work based on explicit instructions that have been manually entered by humans. In simpler systems these rules can then be trivially investigated for potential bias: if some part of the system makes a decision based on a protected attribute, such as gender, it could be considered immediately suspect. This kind of surface-level inspection would reveal trivial cases of bias, such as where a human programmer has encoded in the system that election results pertaining to male candidates are more interesting than similar results pertaining to female candidates.

However, such clear-cut examples are, we hope, rare. We believe it is much more likely that the system incorporates some heuristic that reflect unconscious underlying biases, with unintended results. This becomes increasingly probable as the system complexity and the amount of automated data analysis conducted by the system increase. For example, a system producing news about the local housing market might use the average housing prices of an area as part of its decision making about which areas to discuss in the produced news text, assuming a higher price equates to higher newsworthiness. These housing prices, however, are likely to be well correlated with socioeconomical factors of the area population, resulting in coverage that is biased against populations of lower socioeconomical status as a result of not discussing aspects of the housing market relevant to them.

An even more nuanced example of the same phenomena could be observed if the decisions on what areas to report on were based on the absolute change in the housing prices; if the housing prices changed everywhere by the same percentage, the more well-off areas would see significantly higher absolute changes, which in the case of our hypothetical system would result in the same effect as above. As such, the investigation cannot be limited to only protected attributes, but rather all attributes that correlate with protected attributes must also be inspected.

3.2. Bias in Machine Learning Systems

The other major archetype of systems for automated journalism is presented by systems that employ machine learning. These systems differ from the rule-based systems by the fact that their decision-making is not based on human-written rules and heuristics, but rather on rules identified from training data. Most commonly, in supervised machine learning, this training data takes the form of pairs of 'given this input, the system should produce this output,' such as news texts previously written by human journalists paired with the data that underlies each text. While some works have been published on unsupervised text generation methods where the data is not aligned in this way (e.g., Freitag & Roy, 2018), to our knowledge such systems are still rare and suffer from severe limitations in terms of their applicability to automated journalism. A detailed description of unsupervised automated journalism is thus skipped.

In machine learning systems (see e.g., Flach, 2012, for an introduction to machine learning), the human programmers do not explicitly provide the actual rules of processing, but rather provide a framework and a set of assumptions. For example, in the case of a system for producing automatic textual reports of election results, a programmer might make the assumption that the journalistic process being replicated is, effectively, a 'translation' from the numerical results released by the election organizers to the natural language news report. As such, they might elect to implement a neural machine learning model similar in architecture to those used in machine translation, and train it by using examples of 'given this result data, the system should output this textual description.'

The machine learning process then identifies a specific model (analogous to the ruleset developed by-hand above) that minimizes the average difference between what the model outputs for an input in the training dataset and what the expected output was. In other words, the training attempts to identify a process that mimics the process that generated the training samples as closely as it is able. The degree to which this process succeeds is still limited by factors such as the amount of training data (it is hard to learn things of which there are no examples) and the model architecture (the learned model is restricted by the architecture selected by the human developer, and a badly selected architecture might be fundamentally unable to mimic the process that generated the training data).

Another issue is presented by overfitting, where the learned model might incorporate assumptions that hold for the training data but do not generalize to other cases. Even state-of-the-art machine learning systems for natural language generation suffer from this type of behavior in what is referred to as 'hallucination' in the technical literature. That is, they produce output not grounded in the input data, but based solely on strong correlations found in the training data. Such behavior has been identified in state-of-the-art systems in various domains, ranging from very constrained restaurant description tasks (Dušek, Novikova, & Rieser, 2020) to sports news generation (Puduppully, Dong, & Lapata, 2019).

When discussing bias, the model definition, its architecture, is significantly less important than the examples from which the system is trained. An important aspect of supervised machine learning is that the system truly does its best to mimic a process that could have generated the training data it observes. This means that even if the programmer allowed the system to consider some protected attributes, such as gender, the system only does so if the behavior in the training data seems to be influenced by

said attributes. This, however, also means that if there are any biases in the training data, these are also learnt. This applies whether the biases are intentional or not.

At the same time, however, simply removing a potentially biasing variable from the input is insufficient to ensure that the system does not act in a biased manner and many 'debiasing' techniques can simply hide the issue without solving it (e.g., Gonen & Goldberg, 2019; Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018). As long as the underlying bias exists in the training data, even if the identified variable causing the bias is removed, the system will locate so called proxy variables to encode said bias into the model (Kilbertus et al., 2017). For example, if the training data for a system making loan decision was provided by humans that were discriminatory by providing smaller loans to non-white applicants, a sufficiently complex model might learn to observe whether the name of the applicant or their postal code is indicative of a high likelihood of being non-white as a proxy for the race of the applicant, even if the race was not explicitly provided as input to the system. In the context of automated journalism, a machine learning system would thus learn any biases present in the news stories that were used to train it. As such, the 'unbiased' algorithm would simply be faithfully replicating and entrenching any pre-existing biases in the news text used to train it.

4. Detecting Bias in and with Automated Journalism

As for detecting bias in systems for automated journalism, we see three primary scenarios where such an investigation could be undertaken: a scenario of a clear box system, a scenario of a cooperative operator with a black box system, and a scenario where only system outputs are available. We next discuss each in turn, considering how the system might be diagnosed for bias given the constraints of the scenario.

4.1. Full Transparency

Clear box investigations depend on the ability to inspect the internal workings of the automated journalism system. As such, they are only possible in cases where the operator of the system is cooperative, allowing access to the source code of the system. Furthermore, they are in practice limited to rule-based systems: even if a modern machine learning model was made available to experts, the systems tend to be so immensely complex that they are, in practice, black boxes.

Given access to a rule-based system, it should be in principle possible to investigate the logic and the rules employed by the system and determine whether any of them are blatantly biased. For example, as noted previously, any rules where the system directly considers a variable related to, for example, gender, is immediately suspect of introducing gender bias into the report and can be investigated further. Such an investigation, however, becomes increasingly difficult when one attempts to identify nuanced effects such as those described in the housing price report example shown before.

To identify more nuanced (potential) biases and to investigate systems that are too complex for manual inspection of the system's internal workings, a method based on system input variation might be more practical. Notably, this method still requires some level of cooperation from the system operator but does not require access to the system internals, and as such is also applicable to black box systems. In this process, samples of slightly varied system inputs are prepared, and ran through the system in sequence and the results inspected for differences.

4.2. Cooperative Operator with a Black Box System

An example of such a cooperative black box case would be a machine learning system producing reports of election results. In such a case, one can take the election results that act as the system's input and produce a variation of those results where potentially bias-inducing variables are modified. For example, the researcher could produce a copy of the system input where all the genders of the candidates have been changed but the input is otherwise left as-is. Producing output from both the unmodified and modified inputs would then allow for a comparison of the output texts, so that any differences can be inspected for potential bias. Continuing the example, observing changes between the two datasets in, for example, the order the results are discussed in would give rise to suspicion of potentially biased treatment of the different genders. In fact, any significant changes in lexical and content selection should be investigated in detail.

4.3. Output Only

In cases where the system operator is not cooperating the investigation must be conducted solely based on the available system outputs. From the point of view of the applicable methods, this case is indistinguishable from the case of a researcher conducting an analysis of human-written news, with the potential exception of significantly higher amount of texts available for analysis. We hypothesize, that in this case the role of automated journalism can be reversed, so that automated journalism can help highlight bias in news texts, whether produced by humans or computers.

A relatively simple method for natural language generation is provided by language models. In general terms, a language model is a machine learning model that describes how likely a sentence is based on training data the model was trained on. Consequently, many language models can be used to generate language by querying the model for 'what is the most likely next word, if the preceding words are….' Due to their simplicity, they are currently not very useful for generating real news, even if they do have other applications in the field of natural

language processing. At the same time, if trained on a large collection of news articles, they in effect learn what an 'archetypical' news article looks like and can mimic that style.

Previous technical works, such as those by Sheng, Chang, Natarajan, and Peng (2019), have demonstrated how language models can be interrogated for bias. In their experiment, they construct pairs of sentence starts, such as 'the woman worked as/the man worked as,' and completed the sentences using a language model. Their analysis of the sentence completions revealed the language model had internalized deep societal biases and reflected them in its output.

While standard language models are not suitable for automated generation of real news, we hypothesize that a language model trained on a sufficient amount of training data produced by a news automation system would learn and retain all the biases of the original system, in effect functioning as a proxy. The language model could then be interrogated for bias, for example using the method of Sheng et al. (2019), and any evidence of bias in the language model would be indicative of a potential bias in the underlying system.

While the wide variety of methods for language modelling are too numerous to enumerate here, it is notable that the most recent advances in language modelling take advantage of word embeddings (e.g., Bengio, Ducharme, Vincent, & Jauvin, 2003; more recently, Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018). In word embeddings (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), words (or sometimes subword-units) are represented as points in a multidimensional vector space. Due to the way the word embedding model is trained, these spaces have several intriguing properties, a principal one being that words that are used in similar contexts in the observed texts are located close to each other in the vector space. Therefore, the nearness of two words in this vector space approximates the semantic relatedness of the words. This same mechanism, however, means that word embeddings trained on a text corpus internalize biases from said corpus (e.g., Gonen & Goldberg, 2019). This has two important consequences.

First, when training a language model as suggested above, care must be taken to ensure that bias is not introduced into the language model via use of word embeddings pretrained on another corpus. Consider, for example, a situation where a language model trained on news texts shows potential bias. If the language model is based on word embeddings pretrained on a highly biased corpus, it would not be clear to what degree the observed biases were incorporated into the model from the news text and to what degree from the biased word embeddings. This problem can be avoided by either using a language model that is not based on word embeddings, or preferably by training both the language model and the word embeddings from scratch. While this procedure prohibits taking advantage of the state-of-the-art pretrained language models such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), it should ensure that any biases observed in the final model come from the texts being inspected.

Second, the tendency of word embeddings to internalize biases also present an opportunity. Previous works (e.g., Caliskan et al., 2017; Knoche et al., 2019) have trained word embeddings from various textual corpora in order to detect biases in said texts. For example, given a word embedding model trained on a newspaper corpus, it is possible to inspect whether keywords indicating either a positive or negative affect are, on average, close to the word 'white' than to the word 'black.' A situation where positive keywords are on average closer to the word 'white' than to 'black' indicates that the corpus contains potential racial biases.

Notably, neither of these last two methods (training and inspecting either language models or word embeddings) is in any way dependent on the data underlying the model being derived from a news generation system. Rather, they could be applied to all kinds of news texts, including those produced by human journalists. Similarly, these latter methods might be useful even in scenarios where the system operator is cooperating. As noted by Diakopoulos (2015), reverse engineering can "tease out consequences that might not be apparent even if you spoke directly to the designers of the algorithm" (p. 404). Indeed, it seems unlikely that a rule-based system for automated journalism would be biased on purpose, and more likely that any potential biases are subtle and introduced unintentionally.

5. Conclusions

In this work, we have briefly described what automated journalism is, including a description of the two archetypical technical methods to conduct news automation: rule-based and based on machine learning. We have identified two major categories of bias that can appear in the output of such systems: content bias and language bias. We then provided a description of the mechanisms that might result in biased output from systems for automated journalism, as well as mechanisms through which these biases could be identified. An important observation is that while the mechanisms require an underlying human source for the bias, the biases can emerge in the system without human intention and in very subtle manners.

Our investigation of bias in automated journalism highlights that automatically produced text needs to be inspected for bias just as human-written texts do. The applicable methods, however, depend on the level of cooperation from the system operator as well as the technical details of the system. In cooperative cases more rigorous inspections of automated systems are possible, yet in some cases the investigation is not meaningfully distinguishable from an investigation of human-written texts. As a result, we note that methods such as the one proposed above could also be applied to investigating the biases of human-written news.

We observed that the belief of unbiased automated journalism seems to predicated on two assumptions: that removing the individual—or the first level of hierarchy of influences (Reese & Shoemaker, 2016)—is sufficient to remove bias, and that using automation indeed removes the individual's effect.

Starting with the second assumption, our investigation above indicates that while automation can obscure the influence of the individual, which would naturally lead to assumptions such as above, automation does not remove the influence of the individual. In case of a rule-based system, the individuals who influence the output are those who build the system and decide what rules it should follow. In case of machine learning, the individual is further removed but still has immense effects on the system's actions through their role as a producer and selector of the training data. In either case, the individual remains, albeit obscured by the system itself.

As for the other assumption, that removing the individual removes bias, we point to the fact that this assumption ignores the possibility of influences imposed by the higher levels of Reese and Shoemaker's (2016) hierarchy. In other words, the belief that the removal of the individual removes bias is predicated on the assumption that bias is created by the individual. Such beliefs overlook societal and organizational biases and the nature of the organization and the society as a collective of individuals.

It warrants repeating that automated journalism fundamentally requires an individual or a collective of individuals to define (whether explicitly through programming rules or implicitly by producing and selecting the training data that tells the system what to do) a set of frames through which the data underlying the story is portrayed (e.g., Entman, 1993; Scheufele, 1999). Any claim of the resulting system being 'unbiased' implicitly insinuates that the frames employed are also unbiased, or alternatively overlooks their existence in the first place. Unless these frames are highlighted and scrutinized—both in academia and outside of it—they risk being entrenched and becoming axiomatic. It is for this reason that investigating automated journalism for bias is so important: By obscuring the individual, automation risks obscuring the framing, hiding both the underlying individual and structural biases. This also has consequences for researchers investigating automated journalism for bias: significant care must be taken to identify origins, originators and contexts of any identified biases. For example, the use of machine learning does not preclude a bias originating from a specific individual.

We believe future work needs to be undertaken on at least two fronts. First, computational methods for identifying bias should be extensively trialed in terms of applicability to the analysis of journalistic texts, with the aim of producing a clear description of usefulness and usability, especially to those without extensive technical knowledge. Optimally, the work should lead to easy-to-use tools for both technical and non-technical researchers. Second, the methods for user-cooperative scenarios need to be tested in detail on real-world systems for automated journalism to determine best practices for conducting such audits, and for determining the origins of the discovered biases.

Automated journalism raises a multitude of ethical questions without obvious answers. For example, attributing authorship of computer-generated texts is a difficult task (Henrickson, 2018; Montal & Reich, 2017), which in turn raises the question of credit, and responsibility, for the end product. It is our opinion that automated journalism cannot be allowed to become a smoke screen for eluding responsibility. In terms of practical recommendations, we point the reader towards the succinct but well thought out guidelines published by the Council for Mass Media in Finland (2019). In short, we concur with the view that automated journalism is a journalistic product, hence the control and responsibility must always reside with the newsroom, ultimately in the hands of the editor in chief. In order to ensure that editors can take this responsibility, developers of automated journalism are liable for creating systems that are transparent and understandable, with auditing providing one way of achieving this goal.

Acknowledgments

This article is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The second author's work was enabled by a personal grant from The Media Industry Research Foundation of Finland and C. V. Åkerlund Media Foundation.

Conflict of Interests

The corresponding author is employed in a joint research project with various European media companies.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. Retrieved from https://www.propublica.org

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Caswell, D., & Dörr, K. (2018). Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice*, *12*(4), 477–496.

Council for Mass Media in Finland. (2019). Statement on marking news automation and personaliza-

tion. *Council for Mass Media in Finland*. Retrieved from http://www.jsn.fi/en/lausumat/statement-on-marking-news-automation-and-personalization

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human Language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Stroudsburg, PA: Association for Computational Linguistics.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415.

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Cambridge, MA: Harvard University Press.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809–828.

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, *4*(6), 700–722.

Dušek, O., Novikova, J., & Rieser, V. (2020). Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, *59*, 123–156.

Eastman, S. T., & Billings, A. C. (2000). Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, *24*(2), 192–213.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58.

Fischer-Hwang, I., Grosz, D., Hu, X. E., Karthik, A., & Yang, V. (2020). *Disarming loaded words: Addressing gender bias in political reporting*. Paper presented at Computation + Journalism '20 Conference, Boston, MA.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Freitag, M., & Roy, S. (2018). Unsupervised natural language generation with denoising autoencoders. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3922–3929). Stroudsburg, PA: Association for Computational Linguistics.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *Cornell University*. Retrieved from https://arxiv.org/abs/1609.07236

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 609–614). Stroudsburg, PA: Association for Computational Linguistics.

Graefe, A. (2016). *Guide to automated journalism*. New York, NY: Tow Center for Digital Journalism.

Hackett, R. A. (1984). Decline of a paradigm? Bias and objectivity in news media studies. *Critical Studies in Media Communication*, *1*(3), 229–259.

Harcup, T., & O'Neill, D. (2016). What is news? *Journalism Studies*, *18*, 1470–1488.

Helsingin Sanomat. (2018, March 3). *Helsingin Sanomat lisää naisten osuutta artikkeleissaan* [*Helsingin Sanomat* increases the share of women in its articles] [Press release]. Retrieved from https://sanoma.fi/tiedote/helsingin-sanomat-lisaa-naisten-osuutta-artikkeleissaan

Henrickson, L. (2018). Tool vs. agent: Attributing agency to natural language generation systems. *Digital Creativity*, 29(2/3), 182–190.

Hofstetter, C. R., & Buss, T. F. (1978). Bias in television news coverage of political events: A methodological analysis. *Journal of Broadcasting & Electronic Media*, *22*(4), 517–530.

Hooghe, M., Jacobs, L., & Claes, E. (2015). Enduring gender bias in reporting on political elite positions: Media coverage of female MPs in Belgian news broadcasts (2003–2011). *The International Journal of Press/Politics*, *20*(4), 395–414.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 656–666). Red Hook, NY: Curran Associates.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In W. R. Johnson & K. Markel (Eds.), *AEA papers and proceedings* (Vol. 108, pp. 22–27). Nashville, TN: American Economic Association.

Knoche, M., Popović, R., Lemmerich, F., & Strohmaier, M. (2019). Identifying biases in politically biased wikis through word embeddings. In C. Atzenbeck, J. Rubart, D. E. Millard, & Y. Yesilada (Eds.), *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 253–257). New York, NY: Association for Computing Machinery.

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, *13*(7), 759–780.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Red Hook, NY: Curran Associates.

Mindich, D. T. (2000). *Just the facts: How "objectivity" came to define American journalism*. New York, NY: New York University Press.

Montal, T., & Reich, Z. (2017). I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism. *Digital journalism*, *5*(7), 829–849.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). Stroudsburg, PA: Association for Computational Linguistics.

Plattner, T., & Orel, D. (2019). *Addressing micro-audiences at scale*. Paper presented at the Computation+Journalism Symposium 2019, Miami, FL.

Puduppully, R., Dong, L., & Lapata, M. (2019). Data-to-text generation with content selection and planning. In P. Stone, P. Van Hentenryck, & Z. Zhou (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 6908–6915). Palo Alto, CA: AAAI Press.

Reese, S. D., & Shoemaker, P. J. (2016). A media sociology for the networked public sphere: The hierarchy of influences model. *Mass Communication and Society*, *19*(4), 389–410.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge University Press.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, *49*(1), 103–122.

Schwartz, J. (2011). Whose voices are heard? Gender, sexual orientation, and newspaper sources. *Sex Roles*, *64*(3/4), 265–275.

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In d. boyd, J. Morgenstern, A. Chouldechova, & F. Diaz (Eds.), *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68). New York, NY: Association for Computing Machinery.

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3407–3412). Stroudsburg, PA: Association for Computational Linguistics.

Sirén-Heikel, S., Leppänen, L., Lindén, C. G., & Bäck, A. (2019). Unboxing news automation: Exploring imagined affordances of automation in news journalism. *Nordic Journal of Media Studies*, *1*(1), 47–66.

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14). New York, NY: Association for Computing Machinery.

Yleisradio. (2018). Avoin-voitto source code. *Github*. Retrieved from https://github.com/Yleisradio/avoin-voitto

Yu, R. (2014, June 30). How robots will write earnings stories for the AP. *USA Today*. Retrieved from https://eu.usatoday.com/story/money/business/2014/06/30/ap-automated-stories/11799077

About the Authors

Leo Leppänen is a Computer Science PhD Student at the University of Helsinki. He has a MSc in Computer Science and a BA in Language Technology. His research focus is on automated generation of natural language, especially on the generation of factual content such as news and other reports from structured data. He is currently exploring news automation for less-resources European languages.

Hanna Tuulonen is a PhD Student at the Swedish School of Social Science, University of Helsinki. In her PhD dissertation, Tuulonen researches news automation in China, and how Chinese news automation and data-driven media content affect European media practices and content. In 2017, Tuulonen did her MA thesis in Finnish and Swedish news automation practices, and in 2018 she also participated in the Immersive Automation research project (http://immersiveautomation.com).

COGITATIO

Stefanie Sirén-Heikel is a PhD Student in Media and Communication Studies at the University of Helsinki. She has a B.Soc.Sc. in journalism studies and a M.Soc.Sc. in media and global communication from the University of Helsinki. Her research is focused on how algorithmic decision-making and automation affects journalistic values, how journalism is defined, and performed. She has particularly interest in the sociotechnical aspects of newsroom innovation and management. She has a background in broadcast, print and digital journalism.