Measuring Gender Bias in Word Embeddings for Russian Language

Pestova A.S.
Higher School of Economics
Saint-Petersburg, Russia
alpestova1818@gmail.com

Abstract

The problem of gender bias in Natural Language Processing (NLP) models has become a growing concern in the NLP community in recent years. Due to the fact that the texts on which models are trained often contain stereotypes and prejudices, different types of NLP models, regardless of the task and learning algorithms, demonstrate social biases in terms of gender, race, and religion. Word embeddings (WE) as a very common framework in NLP were shown to reproduce various prejudices as well and gender bias, in particular. Existing research on gender bias in word embeddings often focus on English language models and there is no such research for WE for Russian. In this work, word embeddings for Russian language were analyzed in terms of gender bias for the first time. Using Word Embedding Association Test method and an extended list of analyzed word categories, it was shown that Russian language word embeddings preserve and reproduce gender bias in various topics.

Keywords: gender bias, bias, word embeddings, Natural Language Processing, Russian language
DOI: 10.28995/2075-7182-2021-20-1151-1160

Измерение гендерной предвзятости русскоязычных векторных представлений слов

Пестова А.С.
Higher School of Economics
Санкт-Петербург, Россия
alpestova1818@gmail.com

Аннотация

Проблема гендерной предвзятости моделей автоматической обработки естественного языка (NLP) все больше беспокоит сообщество NLP в последние годы. Из-за того, что тексты, на которых обучаются модели, часто содержат стереотипы и предрассудки, разные типы моделей NLP, независимо от задачи и алгоритмов обучения, демонстрируют социальные предубеждения с точки зрения гендера, расы и религии. В предыдущих исследованиях было показано, что модели векторных представлений слов также воспроизводят различные предрассудки, в том числе, и гендерные. Существующие исследования гендерной предвзятости моделей векторных представлений слов анализируют, в основном, модели для английского языка, а для русского языка подобного анализа не проводилось. В данной работе неконтекстуализированные модели векторных представлений слов для русского языка впервые были проанализированы с точки зрения гендерной предвзятости. С помощью метода WEAT и расширенного списка категорий слов было показано, что русскоязычные модели сохраняют и воспроизводят гендерные предубеждения в различных темах.

Ключевые слова: гендерная предвзятость, векторые представления слов, русский язык

1 Introduction

The problem of fairness of algorithms and social biases contained in them has become a growing concern for ML/AI community and led to active research in this area [7, 32]. The source of concern is that machine learning models can learn and reproduce bias presented in the training data [20, 5].

This is also true for Natural Language Processing (NLP) models that are trained on various text corpora and reproduce bias presented in texts. Different types of models, regardless of the task and training methods, demonstrate social biases in terms of gender, race, and religion [24, 36]. In particular, the problem of gender bias in NLP models has become a growing concern in the NLP community in recent years [27, 28].

NLP models are demonstrated to preserve and reproduce gender bias. For instance, [26] demonstrated that Google Translate contains gender prejudices. Male defaults in its translation are salient and exaggerated especially when translating texts mentioning occupations in fields like STEM that are stereotypically more male. Gender bias in terms of occupation was also found in language models [38], sentiment analysis models [4].

Word embeddings (WE) is a very common framework in NLP allowing to represent words and phrases as vectors in a multidimensional space [9]. A distinctive feature of WE is that vectors of the semantically similar words are close together. The difference between words' vector representations can show meaningful semantic relationships between these words which is also known as word analogies. It was shown that word embeddings can capture different social biases as well and gender bias, in particular [21, 6, 37, 14]. Several methods have been proposed for measuring gender bias in word embeddings in previous studies [21, 6, 37, 14]. Most research analyze English language word embeddings, but there are papers which study other languages, for instance Spanish and French [17, 12], German [17], Dutch [23, 34].

However, there is no research on social biases in word embeddings for the Russian language. The aim of this work is to study whether gender bias is present in different Russian-language word embeddings models and in what topics. WE for Russian were analyzed with the Word Embedding Association Test method [6] in terms of gender bias in 7 categories: career vs family, math vs arts, science vs arts, intelligence vs appearance, physical vs emotional strength, STEM vs humanities, rationality vs emotionality. Depending on the model type and corpus, Russian-language word embeddings are shown to contain gender bias to a varying degree.

2 Related Work

2.1 Gender bias in Texts

Fiction, news, texts crawled from the Web, Wikipedia are often used as training corpora for NLP models. These texts are demonstrated to be prejudiced in terms of gender. For instance, studies of children's fiction showed that women and men are portrayed from the point of view of their traditional roles. Males are shown as strong people having better jobs, less involved in household chores and childcare, while females keep the house doing chores, are often helpless, probably do not work or have less prestigious job [1, 29]. Women are also generally less mentioned and described in fiction than men [33].

In news media, females are also less visible than males [35, 30, 13], appear in stereotypically female sections such as people, culture and society [25]. They are mostly mentioned in topics of fashion and beauty contests, family relationships and childbirth [35], and are often portrayed sexually or with reference to their traditional gender roles such as mother and wife [2].

Wikipedia, which is a popular source of training data in NLP, is demonstrated to be biased as well. Women's pages on Wikipedia contain information about their romantic and family relationships, marriages and divorces more often than men's pages [3, 16]. What is more, Wikipedia underrepresents women in stereotypically both female and male occupations [13].

2.2 Measuring Gender Bias in Word Embeddings

In terms of research on fairness of machine learning and neural network models, bias is defined as any unfair regularities in training data and thus in the models themselves. From the decision-making perspective, fairness is the absence of any favoritism or preconception towards a person or a group of people [11]. So, we could say that the algorithm is unfair and biased if its solutions are skewed toward some group of people.

The first method for measuring gender bias in word embeddings is proposed by [21]. This method is based on identifying a gender subspace using gender-specific words vector representations and measuring the cosine of angle between gender-neutral words (for instance, different occupations) and the gender subspace. The sign of cosine shows in which direction a word is biased (to male or female direction), and the absolute value identifies a degree of bias. The total bias is then measured as the average of all the absolute values of bias for each word. The authors demonstrate word embeddings trained on Google News to be biased.

Another method proposed by [6] is based on the idea of Implicit Association Test (IAT) [15] which is mostly used for measuring stereotypes held by people, for example, associating female names with stereotypically female characteristics. In the IAT, people are suggested to pair two concepts (two set of words) that seem related and similar to them, as opposed to two concepts that seem different for them. The response times of answers are measured and then used for comparing people associations between concepts - longer response times correspond to less association. Similar to the IAT, [6] propose the Word Embedding Association Test (WEAT) which is a statistical test that measures bias in word embeddings between two sets of target words and two sets of attribute words but, instead of reaction time in the IAT, it uses cosine similarities between words' vector representations. As for gender bias, the authors show the bias in the following categories: career vs. family activities, math vs. arts and science vs. arts. It was demonstrated that female names are less associated with career words and more with family words if compared to male names. What is more, female terms are also more related to art rather than science and mathematics.

The WEAT has become a popular method for measuring social biases in word embeddings [8, 17, 12]. It has been shown that it is suitable for the analysis of WE for languages with grammatical gender [17, 12].

Confirmed presence of gender bias in WE inspired research on methods of mitigating bias from WE models. [21] suggest the algorithm that can decrease the gender bias in WE as a post-processing step. [19] propose another method to debias vectors during training the model. However, [14] argue that debiasing algorithms used in the previous research cannot overcome the problem of gender bias in the word embeddings as there still remains the indirect bias that is reflected in the distances between gender-neutral words even after debiasing their vector representations. Another method is creating gender-balanced corpora for training the models [22], which is shown to still preserve gender bias in some categories [8].

3 Data and Methods

3.1 Choice of Word Embeddings

. For the analysis in this paper, Russian-language word embeddings were taken from the RusVectores[1] website [18] where pre-trained embeddings for Russian are uploaded for free download and use. Models were chosen in such a way that they were trained on different corpora for the sake of model comparison and representativeness. Thus, the following models were used for the analysis (for more convenient reference to the models in the following sections, they have been given short names):

a) **RNC_cbow** (ruscorpora_upos_cbow_300_20_2019): Word2Vec CBoW embeddings trained on Russian National Corpus

Russian National Corpus[2] consists of contemporary fiction, modern drama, memoir and biographical literature, journalism and literary criticism, news, scientific, popular science and educational texts, religious texts, technical texts, official business and legal texts, everyday texts. The share of literary texts (including drama and memoir) is no more than 40% .

b) **RNC-Wiki_skip** (ruwikiruscorpora_upos_skipgram_300_2_2019): Word2Vec SGNS embeddings trained on Russian National Corpus and Wikipedia

In addition to the Russian National Corpus, in this WE, dump of Russian Wikipedia for 2019 is used.

[1] `https://rusvectores.org/ru/models/`

[2] `https://ruscorpora.ru/new/corpora-structure.html`

c) **Tayga_skip** (tayga_upos_skipgram_300_2_2019): Word2Vec SGNS embeddings trained on the webcorpus Tayga[3] [31]

This corpus consists of literary texts, social media, subtitles, news, poems and other texts. The subcorpus of poems was not used for training this WE, so the literary texts make up 95% of the used corpora .

d) **News_skip** (news_upos_skipgram_300_5_2019): Word2Vec SGNS embeddings trained on Russian language news

e) **GeoWAC_fast** (geowac_lemmas_none_fasttextskipgram_300_5_2020): FastText CBoW embeddings trained on the corpus GeoWAC [10]

GeoWAC is the geographically balanced corpus with texts from Common Crawl[4] project which is an open repository of web crawl data. For these WE, a sample of Russian-language documents from the corpus was used.

3.2 Word Embeddings Association Test

The WEAT method which was introduced by [6] for measuring bias in word embeddings is used for analysis of gender bias as it was shown that it is suitable for the languages with grammatical gender (Russian is also a language with grammatical gender). The null hypothesis is that the two sets of target words which we suspect to be biased are not different regarding their relative similarity to the two sets of attribute words (male and female terms).

In formal terms, there are two sets of target words of equal size X and Y, and two sets of attribute words A and B. The cosine of the angle between two vectors $\vec{a}$ and $\vec{b}$ is denoted as $cos(\vec{a},\ \vec{b})$. Then, the test statistic is calculated as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{1}$$

where $s(w, A, B)$ is the measure of association between the target word w and two attribute sets A and B:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \tag{2}$$

So, $s(w, A, B)$ measures the difference in similarities of the target word w with the words in attribute sets. And $s(X, Y, A, B)$ calculates the differential association of the sets of target words with the attribute words. For example, if a positive value is obtained, it means that the set X is more associated with the set A then with the set B, compared to the set Y. In contrary, if a value is negative, then the set X is more associated with the set B then with the set A if comparing to the set Y.

In the original paper by [6], the permutation test is used for measuring the likelihood of the null hypothesis. In other words, the authors calculate the probability that the observed or greater difference in sample means would be obtained by a random permutation of target words. In formal terms, $\{(X_i; Y_i)\}_i$ is all the partitions of $X \cup Y$ into two sets of equal size. Then, the one-sided p-value of the permutation test is calculated as:

$$Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)] \tag{3}$$

Following other similar studies [8, 17], in this paper, a randomization test with 100000 iterations was used because a full permutation test can quickly become computationally intractable. A word category is considered to be significantly biased if p-value is less than 0.05. The effect size is calculated as Cohen's d [6]:

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std} - \text{dev}_{w \in X \cup Y} s(w, A, B)} \tag{4}$$

[3] https://tatianashavrina.github.io/taiga_site/

[4] http://commoncrawl.org/

The code for computations of gender bias for this study is available here [5].

3.3 Word Categories for Analysis

Two lists of male and female terms are used in this paper, following [6, 8, 17]. As for the target words, seven pairs of word sets were chosen in order to measure whether there is difference in gender bias between two lists. Sets of words representing the topics of **career and family**, **math and arts**, **science and arts** were taken following [6]. [37] expanded this list with the following topics: **intelligence and appearance**, **physical and emotional strength**. And in the paper [17], the word categories represented **STEM and humanities**, **rationality and emotionality** are additionally analyzed.

In this paper, all the mentioned word lists are taken for the analysis. Words from the previous research were selected and translated into Russian, with the exception of those where the translation was ambiguous and did not reflect the concept of the topic. For instance, all the pronouns were deleted from the sets of male and female terms since they are removed as stop words in the used word embeddings. Words that are not so common in Russian were removed, for example, the word *cousin* in the set of words on the topic of family (direct translation is not commonly used in Russian, and translation / is already a phrase, not a word with its own embedding). Other word lists have been modified in the same way, words for STEM and humanities were greatly changed and new words were added to the lists since, when translated, most words became two-word phrases. All analyzed words were used in the lemmatized form, because the word embeddings that were used contains lemmas. Full lists of words in Russian and their translations into English can be seen in Section 1 in Appendix.

4 Results and Discussion

All the obtained results can be seen in Table 1. Further, the results by word categories will be considered in more detail.

E1: career vs family

The presence of gender bias in the topic of career and family is detected in all the word embeddings except the model GeoWAC_fast. The biggest effect size is for the model Tayga_skip. Thus, we can conclude that in most models female terms are less associated with career words and more with family words if compared to male terms.

E2: math vs arts

Gender bias in terms of association with math/arts topics is found only in the model RNC_cbow. In other words, in these word embeddings, female words are more realted to art rather than math if compared to male words. However, in other models the associations are not statistically significant.

E3: science vs arts

In this category, gender bias was detected in the models RNC_cbow and Tayga_skip. So, female words are more associated with art rather than science in comparison to male words.

E4: intelligence vs appearance

Gender bias in terms of association with the words describing intelligence and appearance is found in all the word embeddings. Moreover, the bias in this category has the largest effect size in almost every model. Thus, we can conclude that male terms are more related to words describing intelligence, while female terms are more related to the topic of appearance.

E5: strength vs weakness

As it can be expected, word category of physical and emotional strength and weakness is biased in terms of gender in almost every model that was analyzed, except the model GeoWAC_fast. Women are more associated with words on the topic of weakness than strength, comparing to males.

E6: STEM vs humanities

[5] https://github.com/Pstva/gender-bias-ru-word-embeddings

This word category is found to be statistically significant only in the models RNC_cbow and Tayga_skip. In these word embeddings, females are less related to STEM and more associated with humanities compared to males.

E7: rationality vs emotionality

In terms of word category for the topic of rationality and emotionality as character traits, the models RNC_cbow, Tayga_skip and News_skip appear to be biased. So, in this models, female terms are more associated with words describing emotionality rather than rationality, if comparing to male terms.

	RNC_cbow		**RNC-Wiki_skip**		**Tayga_skip**	
Word Categories	***d***	***p-value***	***d***	***p-value***	***d***	***p-value***
E1: career vs family	0,262	**0,0201**	0,210	**0,0281**	0,411	**0,0005**
E2: math vs arts	0,588	**0,0159**	0,243	0,1607	0,667	0,1318
E3: science vs arts	0,469	**0,0244**	0,059	0,3822	0,713	**0,0374**
E4: intelligence vs appearance	0,784	**0,0001**	0,735	**0,00002**	0,916	**0,0002**
E5: strength vs weakness	0,455	**0,0189**	0,377	**0,0057**	0,654	**0,0258**
E6: STEM vs humanities	0,441	**0,0346**	0,086	0,3945	0,990	**0,0445**
E7: rationality vs emotionality	0,503	**0,0152**	0,341	0,0546	0,384	**0,0390**

	News_skip		**GeoWAC_fast**	
Word Categories	***d***	***p-value***	***d***	***p-value***
E1: career vs family	0,308	**0,0063**	0,064	0,2662
E2: math vs arts	0,130	0,1397	-0,063	0,6762
E3: science vs arts	0,155	**0,0403**	-0,250	0,8937
E4: intelligence vs appearance	0,314	**0,0021**	0,653	**0,0008**
E5: strength vs weakness	0,324	**0,0111**	0,252	0,1400
E6: STEM vs humanities	0,014	0,4812	0,043	0,4095
E7: rationality vs emotionality	0,703	**0,0022**	0,170	0,2309

Table 1: Results for WEAT hypothesis test for seven word categories and five word embeddings for Russian language. Effect size (Cohen's d) and p-value is reported. Statistically significant gender bias is indicated by the p-values in bold ($p < 0.05$).

All in all, the models **RNC_cbow** and **Tayga_skip** are found to be the most biased in analyzed word categories. Both corpora that models are trained on mostly consists of literary texts, which were demonstrated to contain gender stereotypes.

The model **News_skip** appears to be biased in 5 out of 7 word categories, however, the effect sizes are smaller for almost all the word categories than in the models **RNC_cbow** and **Tayga_skip**. Biases found in the model correspond to the previous research as news are demonstrated to be highly prejudiced in terms of gender.

RNC-Wiki_skip contains gender bias only in 3 out of 7 word categories. It might seem that the explanation for why these embeddings are less biased than the model **RNC_cbow** trained on Russian National Corpus only is that Wikipedia texts contain fewer stereotypes and somehow decrease the bias in these WE. However, these models are built with the different word2vec methods (CBoW and Skip-gram) and with different window sizes which can also influence the presence of bias in embeddings. Larger window sizes capture broader information of words similarities [8] which can be an explanation of reduced bias in the model with smaller window size (it is 5 for the model **RNC-Wiki_skip** and 20 for the model **RNC_cbow**).

The model **GeoWAC_fast** turned out to be the least biased among all analyzed models, as statistically significant bias was found only in one out of seven word categories. Unfortunately, it is impossible to conclude from this analysis whether the fact that the bias was undetected in most categories is an advantage of the corpus GeoWAC or the method fasttext for training embeddings.

5 Conclusion and Future Work

In this study, Russian language word embeddings are demonstrated to preserve and reproduce gender bias in different topics. It was shown that word embeddings trained on corpora of different kind contain gender bias to a varying degree.

However, future research is needed to study the role of corpus composition, hyperparameters (for instance, window size) and model types of word embeddings in preserving gender bias. Moreover, it is necessary to study whether other methods for measuring gender bias are suitable for analysis of word embeddings for Russian language.

6 Appendix

6.1 Lists of words

Here are the lists of attribute words with male and female terms and target words used for each category in the analysis. Words are listed in Russian and their translations into English are given.

Attribute words
Male and female terms:
A: мужчина, мужской, мальчик, брат, сын, отец, папа, дедушка, дядя
(man, male, boy, brother,son, father, father, grandfather,uncle)
B: женщина, женский, девочка, сестра, дочь, мать, мама, бабушка, тетя
(woman,female, girl, sister, daughter, mother, mother, grandmother, aunt)

Target words
E1: career and family
X: руководитель, менеджмент, профессионал, корпорация, зарплата, офис, бизнес, карьера
(executive, management, professional, corporation, salary, office, business, career)
Y: дом, родитель, ребенок, семья, род, брак, свадьба, родственник
(home, parent, children, family, family, marriage, wedding, relative)

E2: math and arts
X: математика, алгебра, геометрия, уравнение, вычисление, число, сложение
(math, algebra, geometry, equation, computation, number, addition)
Y: поэзия, искусство, танец, литература, роман, симфония, драма
(poetry, art, dance, literature, novel, symphony, drama)

E3: science and arts
X: наука, технология, физика, химия, эксперимент, астрономия, исследование
(science, technology, physics, chemistry,experiment, astronomy, research)
Y: поэзия, искусство, танец, литература, роман, симфония, драма
(poetry, art, dance, literature, novel, symphony, drama)

E4: intelligence and appearance
X: развитый, находчивый, любознательный, гениальный, изобретательный,
проницательный, рассудительный, способный, мудрый,
сообразительный, умный, логичный, вдумчивый, творческий
(precocious, resourceful, inquisitive, genius, inventive, astute, judicious, apt, wise, smart, clever, logical, thoughtful, creative)
Y: привлекательный, соблазнительный, роскошный, румяный, пухлый,
чувственный, великолепный, стройный, лысый, красивый, модный, толстый,
слабый, симпатичный
(attractive, alluring, voluptuous, blushing, plump, sensual, gorgeous, slim, bald, beautiful, fashionable, fat, weak, pretty)

E5: strength and weakness

X: сила, сильный, уверенный, доминировать, мощный, громкий, смелый, успешный, лидер, динамичный, победитель
(power, strong, confident, dominant, potent, loud, bold, succeed, leader, dynamic, winner)
Y: слабый, сдаться, робкий, уязвимый, слабость, уступить, застенчивый, проиграть, хрупкий, бояться, неудачник
(weak, surrender, timid, vulnerable, weakness, yield, shy, lose, fragile, afraid, loser)

E6: STEM and humanities
X: электротехника, машиностроение, информатика, программирование, физика
(electrical engineering, mechanical engineering, computer science, programming, physics)
Y: социология, филология, педагогика, психология, лингвистика
(sociology, philology, pedagogy, psychology, linguistics)

E7: rationality and emotionality
X: разум, рациональность, осознание, мышление, знание, рассудительность
(mind, rationality, realization, thinking, knowledge, prudence)
Y: сентиментальность, чувство, эмоция, религиозность, впечатление, настроение
(sentimentality, feeling, emotion, religiosity, impression, mood)

References

[1] Anderson David A., Hamilton Mykol. Gender Role Stereotyping of Parents in Children?s Picture Books: The Invisible Father // Sex Roles. — 2005. — Feb. — Vol. 52, no. 3-4. — P. 145–151. — online; accessed: http://link.springer.com/10.1007/s11199-005-1290-8 (online; accessed: 2021-02-03).

[2] Arslan Bengü, Koca Canan. A Content Analysis of Turkish Daily Newspapers Regarding Sportswomen and Gender Stereotypes // Annals of Leisure Research. — 2007. — Jan. — Vol. 10, no. 3-4. — P. 310–327. — online; accessed: http://www.tandfonline.com/doi/abs/10.1080/11745398.2007.9686769 (online; accessed: 2021-02-02).

[3] Bamman David, Smith Noah A. Unsupervised Discovery of Biographical Structure from Text // Transactions of the Association for Computational Linguistics. — 2014. — Dec. — Vol. 2. — P. 363–376. — online; accessed: https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00189 (online; accessed: 2021-01-31).

[4] Bhaskaran Jayadev, Bhallamudi Isha. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. — 2019.

[5] Buolamwini Joy, Gebru Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Proceedings of the 1st Conference on Fairness, Accountability and Transparency / Ed. by Sorelle A. Friedler, Christo Wilson. — Vol. 81 of Proceedings of Machine Learning Research. — New York, NY, USA : PMLR, 2018. — Feb. — P. 77–91. — Access mode: http://proceedings.mlr.press/v81/buolamwini18a.html.

[6] Caliskan Aylin, Bryson Joanna J, Narayanan Arvind. Semantics derived automatically from language corpora contain human-like biases // Science. — 2017. — Vol. 356, no. 6334. — P. 183–186. — Publisher: American Association for the Advancement of Science.

[7] Caton Simon, Haas Christian. Fairness in Machine Learning: A Survey // arXiv:2010.04053 [cs, stat]. — 2020. — Oct. — arXiv: 2010.04053. Access mode: http://arxiv.org/abs/2010.04053 (online; accessed: 2021-01-24).

[8] Chaloner Kaytlin, Maldonado Alfredo. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories // Proceedings of the First Workshop on Gender Bias in Natural Language Processing. — Florence, Italy : Association for Computational Linguistics, 2019. — P. 25–32. — online; accessed: https://www.aclweb.org/anthology/W19-3804 (online; accessed: 2021-02-21).

[9] Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in Neural Information Processing Systems / Ed. by C. J. C. Burges, L. Bottou, M. Welling et al. — Vol. 26. — Curran Associates, Inc., 2013. — P. 3111–3119. — Access mode: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[10] Dunn Jonathan, Adams Benjamin. Geographically-Balanced Gigaword Corpora for 50 Language Varieties. — 2020.

[11] Ethical Challenges in Data-Driven Dialogue Systems / Peter Henderson, Koustuv Sinha, Nicolas Gontier et al. — 2017.

[12] Examining Gender Bias in Languages with Grammatical Gender / Pei Zhou, Weijia Shi, Jieyu Zhao et al. — 2019. — Pages: 5287.

[13] Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms / Vivek K. Singh, Mary Chayko, Raj Inamdar, Diana Floegel // Journal of the Association for Information Science and Technology. — 2020. — Nov. — Vol. 71, no. 11. — P. 1281–1294. — online; accessed: https://onlinelibrary.wiley.com/doi/10.1002/asi.24335 (online; accessed: 2021-02-02).

[14] Gonen Hila, Goldberg Yoav. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. — 2019.

[15] Greenwald A., McGhee D., Schwartz J. L. Measuring individual differences in implicit cognition: the implicit association test. // Journal of personality and social psychology. — 1998. — Vol. 74 6. — P. 1464–80.

[16] It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia / Claudia Wagner, David Garcia, Mohsen Jadidi, Markus Strohmaier // arXiv:1501.06307 [cs]. — 2015. — Mar. — arXiv: 1501.06307. Access mode: http://arxiv.org/abs/1501.06307 (online; accessed: 2021-01-25).

[17] Kurpicz-Briki Mascha. Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings. — 2020.

[18] Kutuzov Andrey, Kuzmenko Elizaveta. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers / Ed. by Dmitry I. Ignatov, Mikhail Yu. Khachay, Valeri G. Labunets et al. — Cham : Springer International Publishing, 2017. — P. 155–161. — Access mode: http://dx.doi.org/10.1007/978-3-319-52920-2_15.

[19] Learning Gender-Neutral Word Embeddings / Jieyu Zhao, Yichao Zhou, Zeyu Li et al. — 2018.

[20] Angwin J., Larson J., Mattu S., Kirchner L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. — 2016. — Access mode: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (online; accessed: 2020-12-22).

[21] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings / Tolga Bolukbasi, Kai-Wei Chang, James Zou et al. // Proceedings of the 30th International Conference on Neural Information Processing Systems. — NIPS'16. — Red Hook, NY, USA : Curran Associates Inc., 2016. — P. 4356–4364. — event-place: Barcelona, Spain.

[22] Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns / Kellie Webster, Marta Recasens, Vera Axelrod, Jason Baldridge // Transactions of the Association for Computational Linguistics. — 2018. — Vol. 6. — P. 605–617. — Access mode: https://www.aclweb.org/anthology/Q18-1042.

[23] Mulsa Rodrigo Alejandro Chávez, Spanakis Gerasimos. Evaluating Bias In Dutch Word Embeddings // arXiv:2011.00244 [cs]. — 2020. — Nov. — arXiv: 2011.00244. Access mode: http:

//arxiv.org/abs/2011.00244 (online; accessed: 2021-02-24).

[24] Nadeem Moin, Bethke Anna, Reddy Siva. StereoSet: Measuring stereotypical bias in pretrained language models. — 2020. — _eprint: 2004.09456.

[25] Perpetuating Gender Inequality via the Internet? An Analysis of Women's Presence in Spanish Online Newspapers / Ruth Mateos de Cabo, Ricardo Gimeno, Miryam Martínez, Luis López // Sex Roles. — 2014. — Jan. — Vol. 70, no. 1-2. — P. 57–71. — online; accessed: http://link.springer.com/10.1007/s11199-013-0331-y (online; accessed: 2021-02-02).

[26] Prates Marcelo, Avelar Pedro, Lamb Luís. Assessing Gender Bias in Machine Translation – A Case Study with Google Translate. — 2018.

[27] Proceedings of the First Workshop on Gender Bias in Natural Language Processing / Ed. by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, Kellie Webster. — Florence, Italy : Association for Computational Linguistics, 2019. — Access mode: https://www.aclweb.org/anthology/W19-3800.

[28] Proceedings of the Second Workshop on Gender Bias in Natural Language Processing / Ed. by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, Kellie Webster. — Barcelona, Spain (Online) : Association for Computational Linguistics, 2020. — Access mode: https://www.aclweb.org/anthology/2020.gebnlp-1.0.

[29] Pyle Wilma J. Sexism in children's literature // Theory Into Practice. — 1976. — Apr. — Vol. 15, no. 2. — P. 116–119. — online; accessed: http://www.tandfonline.com/doi/abs/10.1080/00405847609542620 (online; accessed: 2021-02-02).

[30] Ross Karen, Carter Cynthia. Women and news: A long and winding road // Media, Culture & Society. — 2011. — Nov. — Vol. 33, no. 8. — P. 1148–1165. — online; accessed: http://journals.sagepub.com/doi/10.1177/0163443711418272 (online; accessed: 2021-01-25).

[31] Shavrina T., Shapovalova O. TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: «TAIGA» SYNTAX TREE CORPUS AND PARSER. — Saint-Petersbourg, 2017.

[32] A Survey on Bias and Fairness in Machine Learning / Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena et al. // arXiv:1908.09635 [cs]. — 2019. — Sep. — arXiv: 1908.09635. Access mode: http://arxiv.org/abs/1908.09635 (online; accessed: 2021-01-24).

[33] Underwood Ted, Bamman David. The Gender Balance of Fiction, 1800-2007. — 2016. — Access mode: https://tedunderwood.com/2016/12/28/the-gender-balance-of-fiction-1800-2007/ (online; accessed: 2021-02-03).

[34] Wevers Melvin. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990 // Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. — Florence, Italy : Association for Computational Linguistics, 2019. — P. 92–97. — online; accessed: https://www.aclweb.org/anthology/W19-4712 (online; accessed: 2021-02-24).

[35] Who Makes The News? Global Media Monitoring Project / S. Macharia, L. Ndangam, M. Saboor et al. — 2015. — P. 34. — Access mode: http://www.media-diversity.org/additional-files/Who_Makes_the_News_-_Global_Media_Monitoring_Project.pdf (online; accessed: 2021-02-01).

[36] The Woman Worked as a Babysitter: On Biases in Language Generation / Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, Nanyun Peng. — 2019. — Pages: 3403.

[37] Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes / Nikhil Garg, Londa Schiebinger, Dan Jurafsky, James Zou // Proceedings of the National Academy of Sciences. — 2018. — Apr. — Vol. 115, no. 16. — P. E3635–E3644. — arXiv: 1711.08412. Access mode: http://arxiv.org/abs/1711.08412 (online; accessed: 2021-02-18).

[38] Yeo Catherine, Chen Alyssa. Defining and Evaluating Fair Natural Language Generation. — 2020.