HOKKAIDO UNIVERSITY

Title	Speciesist language and nonhuman animal bias in English Masked Language Models
Author(s)	Takeshita, Masashi; Rzepka, Rafal; Araki, Kenji
Citation	Information Processing and Management, 59(5), 103050 https://doi.org/10.1016/j.ipm.2022.103050
Issue Date	2022-09
Doc URL	https://hdl.handle.net/2115/92943
Rights	© <2022>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
Rights(URL)	https://creativecommons.org/licenses/by-nc-nd/4.0/
Type	journal article
File Information	ethical_nlp_revise.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Speciesist Language and Nonhuman Animal Bias in English Masked Language Models

Masashi Takeshita[a,*], Rafal Rzepka[b] and Kenji Araki[b]

[a]Graduate School of Information Science and Technology, Hokkaido University
[b]Faculty of Information Science and Technology, Hokkaido University

ARTICLE INFO

Keywords:
Social Bias
Bias Evaluation
Masked Language Model
Animal Ethics
Speciesism

ABSTRACT

Warning: This paper contains examples of offensive language, including insulting or objectifying expressions.

Various existing studies have analyzed what social biases are inherited by NLP models. These biases may directly or indirectly harm people, therefore previous studies have focused only on human attributes. However, until recently no research on social biases in NLP regarding nonhumans existed. In this paper[1], we analyze biases to nonhuman animals, i.e. speciesist bias, inherent in English Masked Language Models such as BERT. We analyzed speciesist bias against 46 animal names using template-based and corpus-extracted sentences containing speciesist (or non-speciesist) language. We found that pre-trained masked language models tend to associate harmful words with nonhuman animals and have a bias toward using speciesist language for some nonhuman animal names. Our code for reproducing the experiments will be made available on GitHub[2].

1. Introduction

Recently, in the field of Natural Language Processing (NLP), Masked Language Models (MLMs) using Transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), widely contributed to the state-of-the-art methods in downstream tasks. However, existing studies suggest that these models inherit social biases (Sun et al., 2019; Blodgett et al., 2020). Here, social bias is defined as the systematic, unequal treatment based on one's social attribute (cf. Stanczak and Augenstein, 2021; Sun et al., 2019). Such biases cause differences in accuracy between majority and minority attributes (e.g. Romanov et al., 2019) and negative generalizations, e.g. in text generation (Liu et al., 2020; Sheng et al., 2019, 2021; Garimella et al., 2021).

The studies of social bias in NLP target gender (e.g. Bolukbasi et al., 2016; Caliskan et al., 2017), race (e.g. Manzini et al., 2019), religion and ethnicity (e.g. Li et al., 2020) and so on, all of which assume human attributes. However, to the best of the authors' knowledge, there are no similar bias studies on **nonhuman animals**.

In this paper, we use templates, corpus-extracted sentences and pre-trained MLMs to investigate if the bias regarding nonhuman, i.e. **speciesist bias**, is inherent in MLMs trained on English corpora.

The bias we investigate in this paper is the **representational bias**, following the classification of Sun et al. (2019) and Blodgett et al. (2020). Currently, nonhuman animals do not use the NLP system directly, so we do not need to consider the idea of, e.g. "performance against the social group of nonhuman animals". On the other hand, we think that we should respect nonhuman animals **for their own sake, not for the sake of humans** (cf. Owe and Baum, 2021), for the reasons described below, and therefore we should study, for example, insulting associations with nonhuman animals and negative stereotyping against them.

The structure of this paper is as follows. This section will be followed by an explanation of the reasons for conducting bias studies on nonhuman animals, in response to possible criticisms. In Section 2, we explain related works – we begin with an overview of social bias research in pre-trained language models, then move on to speciesism and speciesist language. We will describe the experimental setting in Section 3 and the experimental methods in Section 4.

[1]Anonymous previous version of this paper is accessible at https://openreview.net/forum?id=dfqMpjZOgv4, we have also published a paper on this topic in Japanese (Takeshita et al., 2021)

[2]https://github.com/Language-Media-Lab/speciesist-language

*Corresponding author

takeshita.masashi@ist.hokudai.ac.jp (M. Takeshita); rzepka@ist.hokudai.ac.jp (R. Rzepka); araki@ist.hokudai.ac.jp (K. Araki)

ORCID(s): 0000-0001-5853-3262 (M. Takeshita); 0000-0002-8274-0875 (R. Rzepka); 0000-0001-9668-1610 (K. Araki)

In Section 5, we report experimental results. In Section 6 we discuss these results and the limitations of the experiments presented in this paper. In Section 7, given experimental results, we discuss how to mitigate speciesist bias. Finally, we conclude our paper in Section 8.

1.1. Ethical Discussion: Nonhumans and NLP

There may be two possible criticisms of the research objectives of this paper. The first criticism is that there is no ethical problem with the existence of harmful bias to nonhuman animals. However, we should give equal consideration to interests and should not discriminate based on **who** has the interests (Singer, 2015). Even if one does not accept this idea, most people would agree that nonhuman animals deserve some moral consideration (Owe and Baum, 2021)[1]. If this is true, then it is important to study the biases that are harmful to nonhuman animals.

The second potential criticism is that even if nonhuman animals deserve some moral consideration, NLP models with speciesist bias do not harm them because they do not use it directly. However, we think it is important to study the speciesist bias of NLP models for three following reasons.

First, if NLP systems with a speciesist bias are popularized in our society, the bias of the NLP system may affect us and thereby indirectly harm animals (in human animal cases, see Bender et al., 2021)[2]. For example, if an NLP system generates speciesist sentences, the speciesist bias may propagate to readers who read the sentences, and they may acquire an implicit discriminatory bias against nonhuman animals[3]. As we discuss in Section 2.2, most people are already discriminatory against nonhuman animals, but this phenomenon should not be reinforced. In addition, it may be harmful not only to nonhuman animals, but also to those who oppose speciesism and use non-speciesist language. For example, if the text generated is such that nonhuman animals are represented as foods, vegetarians (vegans) who oppose speciesism will be offended and will not use the technology. This will lead to the exclusion of minorities from the technology (cf. Blodgett et al., 2020). If one considers the exclusion from technology of vegetarians or vegans as unimportant, it is discrimination against them (Horta, 2018).

Second, the representational speciesist bias should be considered unwarranted in itself, even if it does no direct harm (Blodgett et al., 2020). The use of language that is insulting to or demeaning nonhuman animals, as described in Section 2.2, is wrong in itself (cf. Hellman, 2008), even if nonhuman animals never recognize the expression.

Third, the biases inherent in word embeddings reflect social biases which exist in our cognition, beliefs and social structures (Caliskan et al., 2017; Garg et al., 2018; Joseph and Morgan, 2020). Therefore, analyzing the speciesist bias in word embeddings and corpora can contribute to research about the influence of this bias on our cognition and society.

For these reasons, we think that it is important to study the speciesist bias in NLP.

2. Related Work

2.1. Social Bias in Pre-trained Language Models

Since the report by Bolukbasi et al. (2016) of an inherent gender bias in word embedding, various social biases have been reported to be inherent in static and contextualized word embedding[4]. Previous studies have typically focused on binary gender and race in bias research (especially, Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019), Nangia et al. (2020), for example, they have created bias assessment datasets for as many as nine different attributes. However, these are all bias studies **only on the human animal**, and to the best of the authors' knowledge, until recently there were **no bias studies on nonhuman animals**.

A notable exception is the very recent study by Hagendorff et al. (2022). They have conducted research on speciesist biases in image recognition, recommendation systems, and analyzed static word embedding (word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)), GPT-3 (Brown et al., 2020) and Delphi (Jiang et al., 2021) by using specific word lists and template sentences. There are two differences between their study and ours. First, our experiments are not limited to a specific word list, and we also conduct analysis without template sentences (Section 4.2). Second, we are analyzing speciesist bias in relation to speciesist language. We hope that this new topic is the beginning of a series of research endeavours for analyzing and mitigating speciesist bias.

[1] You can see the more argument about moral status of about nonhuman animals in Regan (2004), Hagendorff et al. (2022) or textbooks of animal ethics (e.g. Fischer, 2021, ch.2).

[2] "Stochastic parrots" in the title of Bender et al. (2021) is an example of specisist language use.

[3] Menegatti and Rubini (2017) argue that sexist language affects our perception, e.g., adversely has impact on women. We think that speciesist language can have similar effect.

[4] For a survey of bias assessment methods, see Dev et al. (2021b).

Bias assessment methods include intrinsic and extrinsic assessment (Goldfarb-Tarrant et al., 2021). Intrinsic evaluation evaluates the bias in the word embedding space itself, while extrinsic one evaluates the bias in the downstream tasks. Since there are currently no downstream tasks in which nonhuman animals are direct users, we use intrinsic evaluation in this paper. However, we think that extrinsic evaluation is necessary for tasks such as text generation, which is a downstream task but whose effects on unexpected users must be considered.

There are three methods of intrinsic evaluation of bias for contextualised word embeddings, especially MLMs: using **template sentences** (Bartl et al., 2020; Hutchinson et al., 2020; Kurita et al., 2019; May et al., 2019; Tan and Celis, 2019; Webster et al., 2020; Silva et al., 2021; Lauscher et al., 2020; Dev et al., 2020, 2021a), **corpus sentences** (Zhao et al., 2019; Bordia and Bowman, 2019; Basta et al., 2019; Guo and Caliskan, 2020), and **non-template sentences created manually** (Nadeem et al., 2021; Nangia et al., 2020). The method using template sentences allows the analysis to focus on specific aspects of bias and is easy to evaluate once the template sentences are created. In the case of using sentences extracted from a corpus, it is possible to evaluate the bias using actual sentences, which allows for a more realistic experimental setting. The method of using manually created non-template sentences is more costly to create, but allows for evaluation in a wider range of realistic settings than using template sentences, and also helps to evaluate the bias in more appropriately limited settings than using sentences extracted from a corpus.

As we will discuss in detail in Section 4, due to the cost of manually creating sentences and other obstacles inherent in this research, to evaluate speciesist bias we use template sentences and sentences extracted from corpora.

2.2. Speciesism and Language

Speciesism is "the unjustified comparatively worse consideration or treatment of those who do not belong to a certain species." (Horta and Albersmeier, 2020, p.3). Nonhuman animals, as sentient beings, deserve equal consideration with human animals (Singer, 2015, p.40), and we should not discriminate against nonhuman animals. However, we do so, for example by eating their flesh or conducting experiments on them (Singer, 2015, ch.2, 3).

We also treat nonhuman animals as inferior beings or objects in our language use. For instance, "terming a woman a 'dog"' insults all women indirectly and also insults all dogs directly (Dunayer, 1995, p.12). Usual referring to nonhuman animals as "it" or "something," or using "that" or "which" as relative pronouns to indicate nonhuman animals are examples of treating nonhuman animals as objects (Dunayer, 2001, 2003). Dunayer (2001, ch.9) also states that, in the process of slaughtering, people use words such as "harvest", "package" and "process" to hide cruelty. The use of such language and words is referred to as **speciesist language**, in analogy with sexist language and racist language. Non-speciesist language is the use of language that does not treat nonhuman animals as objects or as inferior to human animals, such as using "who" to refer nonhuman animals, instead of using "which".

In addition to research conducted in Animal Ethics field, there are also studies in Corpus Linguistics that analyzed language use regarding nonhuman animals. Jepson (2008) performed discourse analysis on various texts and spoken conversations showing that the word "slaughter" in human context collocates strongly with negative emotions, but lacks such sentiment when used in the context of nonhuman animals. Franklin (2020) also analyzed the use of "killing" terms, such as "kill" and "slaughter", in "People, Products, Pests and Pets" (PPPP)[5] which is an English corpus that contains texts referring to nonhuman animals extracted from various domains such as food-related websites and news articles (Sealey and Pak, 2018).

Existing studies have reported that stylistic biases are reflected in NLP models (Tan et al., 2020; Hovy et al., 2020). Therefore, since the above-mentioned speciesist language and biases in English may be reflected in MLMs, we investigate a possibility of speciesist bias in English MLMs.

3. Experimental setup

The MLMs used in this paper are $\text{BERT}_{\text{LARGE-cased}}$[6], $\text{RoBERTa}_{\text{LARGE}}$[7], $\text{DistilBERT}_{\text{base-cased}}$[8] (Sanh et al., 2019) and $\text{ALBERT}_{\text{large-v2}}$[9] (Lan et al., 2020), which are widely used in current NLP. BERT, DistilBERT and ALBERT were pre-trained on Wikipedia and BookCorpus (Zhu et al., 2015). RoBERTa was pre-trained on Wikipedia, BookCorpus, CC-NEWS, OpenWebText (Gokaslan and Cohen, 2019) and STORIES (Trinh and Le, 2018). DistilBERT and ALBERT are smaller models than BERT and RoBERTa. Model details are shown on Table 7 in Appendix A.

[5] https://animaldiscourse.wordpress.com/
[6] https://huggingface.co/bert-large-cased
[7] https://huggingface.co/roberta-large
[8] https://huggingface.co/distilbert-base-cased
[9] https://huggingface.co/albert-large-v2

Table 1
Animal names used in this research and their frequencies in English Wikipedia. The coloring of animal names was done by the authors: refers to "farm" animals, represents popular nonhuman companions and addresses all remaining species.

Animal name	Frequency	Animal name	Frequency
horse	194,363	deer	43,130
turkey	187,079	seal	42,533
fox	176,569	snake	42,323
human	173,145	persian	39,764
fish	142,508	duck	36,828
dog	127,775	swan	36,556
bird	124,463	sheep	34,433
moth	93,670	chicken	34,231
buffalo	91,392	snail	33,725
robin	89,168	bombay	32,819
cat	83,038	frog	31,922
wolf	78,795	crane	31,328
eagle	78,126	penguin	30,769
bear	69,029	rat	28,851
lion	67,774	monkey	28,144
tiger	60,709	falcon	27,843
beetle	54,887	rabbit	27,039
bat	49,445	beaver	26,421
mouse	48,866	pike	25,392
fly	45,411	pig	25,273
newfoundland	44,353	elephant	24,817
tang	44,245	cow	22,563
butterfly	44,096	molly	21,353

We determine animals we focus on in this paper as follows:

1. We collect animal names from "All Animals A-Z List."[10] We focus on only one-term names.
2. We limited the number of animals for this research by choosing only these which names appear on English Wikipedia[11] more than 20,000 times, resulting in 46 animal names in total.

Our hypothesis is that if MLMs recognize different animals by categorizing them, then similar bias will be found for animals in similar contexts. In this paper, we categorize animals who live in farms to be utilized as flesh marking them in , nonhuman companions in , and other animals in colors, respectively[12]. In Table 1, we show all animal names under investigation, their corresponding colors, and their frequencies in Wikipedia.

As shown in Table 1, the taxonomic classes of the animal names used in this experiment are not the same. For example, "dog" is a taxonomic subspecies, while "bear" is a family. Initially, we attempted to standardize the taxonomic classes, but in doing so, we had to exclude some animal names that are used in everyday life. In the first place, we see no special reason to keep the same taxonomic hierarchy, since the extension of animal names as indicated in everyday terminology may not coincide with the extension of animal names in scientific taxonomy (Yoon, 2009, ch.5, 10). In this experiment, we therefore decided not to standardize the taxonomic classes because this would be undesirable from the viewpoint of bias evaluation.

In addition, some words such as "turkey" and "crane" have senses other than animal names, and it is possible that MLMs express these meanings in their internal parameters. However, we do not exclude such animal names from our experiments because we do not know a priori what meanings and biases are reflected in them.

[10] https://a-z-animals.com/animals/

[11] We use the Wikipedia dataset downloaded on 01/05/2020 from https://huggingface.co/datasets/wikipedia.

[12] Following Crameri et al. (2020), in this paper we use scientific color map (Crameri, 2021) to include people with diverse color vision.

4. Bias Analysis by Speciesist and Non-Speciesist Language

In this section, we explain the method for evaluating bias in MLMs. In this experiment, we want to evaluate the relationship between speciesist language and speciesist bias, thus we assess the speciesist bias of MLMs by changing between speciesist and non-speciesist languages. There are two more reasons for using this method. First, there is no list of biased words for evaluating bias against nonhuman animals, so we need to evaluate the bias in a way that does not rely on a pre-defined word list. Second, while most people are aware of gender or racial bias, we think that only few people consider the speciesist bias to be problematic. Therefore, it would be difficult to hire annotators to create a dataset for speciesist bias evaluation. For example, the word "slaughter" is considered to be a biased word for nonhuman animals. However, since most people would not consider this word to be problematic in the context of meat processing (Jepson, 2008), there is a high probability that annotators would not include this word in the evaluation word list.

In this experiment, we evaluate the bias using two approaches: the template-based and the corpus-based. The template-based approach is commonly used in bias analysis studies of NLP models, and we believe it is useful for discovering specific biases (Kurita et al., 2019; Bartl et al., 2020; Webster et al., 2020). However, this approach may limit the bias that can be evaluated depending on the template prepared (Guo and Caliskan, 2020). For this reason, we use raw sentences extracted from the corpus.

In Section 4.1, we describe the template-based experiment. The template sentences used in this experiment are introduced first, and then two experiments using these sentences are described (Section 4.1.1 and Section 4.1.2, respectively). Finally, in Section 4.2, we explain the corpus-based experiment.

4.1. Template-based Experiment

The basic template sentence we utilize is **"[PRONOUN] is a [ANIMAL] [REL-PRONOUN] is [MASK]."**, where [PRONOUN] slot indicates a pronoun, [ANIMAL] is an animal name, and [REL-PRONOUN] stands for a relative pronoun.

We evaluate bias toward [ANIMAL] by observing the change of predicted probability of words at the [MASK] token by replacing [PRONOUN] and [REL-PRONOUN]. We use the following combinations of [PRONOUN] and [REL-PRONOUN]:

- human-describing sentences (hereinafter referred to as "human sentences")
  - *She* is a [ANIMAL] *who* is [MASK].
  - *He* is a [ANIMAL] *who* is [MASK].
- object-describing sentences (hereinafter referred to as "object sentences")
  - *This* is a [ANIMAL] *which* is [MASK].
  - *That* is a [ANIMAL] *which* is [MASK].
  - *It* is a [ANIMAL] *which* is [MASK].
  - *This* is a [ANIMAL] *that* is [MASK].
  - *That* is a [ANIMAL] *that* is [MASK].
  - *It* is a [ANIMAL] *that* is [MASK].

In human sentences, we use "she", "he", and "who", which generally refer to humans. In object sentences, we use "this", "that", "it", and "which", which are generally used for nonhumans. Since pronouns in object sentences are only in the third person equivalently, only the third person pronouns "she" and "he" are used in human sentences.

Our hypothesis here is that the characteristics of the words that are filled in "[MASK]" will change among animals that are often referred to in the speciesist language and others that are not. For example, not only humans, but also dogs and cats could be referred to by the non-speciesist language, while "farm animals" (e.g. cow and pig) would be addressed by the speciesist language.

4.1.1. Bias Evaluation by Word Probability Differences

We evaluate the bias against animal names using words with a large change rate of average predicted probability between human and object sentences. It is done by averaging predicted probability of the word filled into the [MASK] token in the template sentences. We also investigate the relationship between animals by clustering them using the agreement rate of words with large probability changes. We perform this experiment as follows:

1. Calculating $p^{w_i}_{mean_o}$(name) and $p^{w_i}_{mean_h}$(name), which represent predicted probabilities of words corresponding to [MASK] tokens in each sentence, averaged over object ($p^{w_i}_{mean_o}$) and human ($p^{w_i}_{mean_o}$) sentences, respectively, where name is an animal name and w_i is a token in vocabulary V of the MLM (i.e. $w_i \in V$)
2. Calculating how much this probability changes by $\log \frac{p^{w_i}_{mean_o}(\text{name})}{p^{w_i}_{mean_h}(\text{name})}$
3. Ignoring words w_i if (a) both $p^{w_i}_{mean_o}$(name) and $p^{w_i}_{mean_o}$(name) $< \frac{1}{|V|}$, or (b) insignificant words at the center of the distribution of $\log \frac{p^{w_i}_{mean_o}(\text{name})}{p^{w_i}_{mean_h}(\text{name})}$ for each MLM
4. Calculating Token-Match-Rate (TMR) among animal names
5. Clustering all animals based on TMR with UPGMA algorithm (Michener and Sokal, 1957).

In step 1, we calculate $p^{w_i}_{mean_{o,h}}$ as follows:

$$p^{w_i}_{mean}(\text{name}) = \frac{1}{|T|}\sum_{s\in T}^{|T|} p_{\text{MLM}}(w_i|s(\text{name})) \tag{1}$$

where T is the set of object or human template sentences described above, s(name) is a template sentence filled with an animal name. In step 3, in order to ignore insignificant words, first we calculate z-score of all words w_i of $\log \frac{p^{w_i}_{mean_o}(\text{name})}{p^{w_i}_{mean_h}(\text{name})}$, and then we ignore words with |z-score| lower than a threshold[13]. Here z-score $= \frac{x-\mu}{\sigma}$, x is $\log \frac{p^{w_i}_{mean_o}(\text{name})}{p^{w_i}_{mean_h}(\text{name})}$, μ is the average, σ is the standard deviation. In step 4, where $S^{(i)}$ and $S^{(j)}$ are the obtained sets of words for the i, j-th animal names after step 3, we calculate TMR(i, j) between both sets (cf. Webster et al., 2020; Lauscher et al., 2021):

$$\text{TMR}(i,j) = \frac{|S^{(i)} \cap S^{(j)}|}{\min(|S^{(i)}|, |S^{(j)}|)} \tag{2}$$

In step 5, we cluster animal names by using $1 - \text{TMR}(i, j)$ as distance between i, j-th names.

4.1.2. Bias Evaluation by Sentiment Analysis

In this experiment, we use VADER (Hutto and Gilbert, 2014) for evaluating the sentiment of all words which we obtain from the experiment described in Section 4.1.1. This approach does not take into account context when evaluating sentiment of the words, but we decided to analyze the sentiment of the words themselves, considering the possibility of (non-)speciesist bias in the animal names.

Our hypothesis is that when animals are regarded as objects, they are treated negatively, and therefore more negative words will appear under MASKs in object sentences.

4.2. Corpus-based Experiment

In this section, we explain how the bias is measured in the corpus-based evaluation method. The corpus used in this paper is Books3 (Presser, 2020, see also (Gao et al., 2020)) which totals about 100GB of text and is built only from published books. Thus, it is unlikely to overlap with BookCorpus, which contains unpublished books used for the pre-training of MLMs.

To experiment with corpus-based method, we extract object and human sentences from a given corpus. For the purpose of this research, we extract all corpus sentences that contain relative pronouns referring to animals. We use five relative pronouns: “that”, “which”, “who”, “whose” and “whom”. Our assumption is that these relative pronouns can be used to determine whether (non)human animals are treated as objects or humans in the given sentence.

CoreNLP (Manning et al., 2014) is used to extract sentences containing relative pronouns which refer to an animal name. If the speciesist bias exists in CoreNLP, then there may be a difference in referring precision between human and object sentences. Therefore, we asked a native speaker of English to check whether relative pronouns are correctly referred to an animal name in ten sentences (for each pronoun) randomly extracted from Book3. As a result, one

[13] Here this threshold value was experimentally set to 1.96.

sentence containing "who", and two with "whom" have been marked as incorrect, and all remaining 47 sentences have been judged as having correct references. This suggests that the precision of the parser for the task is relatively high.

For the corpus-based bias evaluation, we replace relative pronouns referring to animal names with [MASK] tokens in extracted sentences. Then, we use MLMs to calculate probabilities of relative pronouns at the [MASK] token. We compare the probabilities for both sets and evaluate the bias as follows:

$$bias = \frac{1}{|H|}\sum_{s_i \in H}^{|H|} \mathbb{1}[p_{object|s_i} > p_{human|s_i}] - \frac{1}{|O|}\sum_{s_j \in O}^{|O|} \mathbb{1}[p_{human|s_j} > p_{object|s_j}] \tag{3}$$

where H and O are the sets of human and object sentences extracted from Books3, and $s_{i,j}$ is a given sentence. $\mathbb{1}[\cdot]$ returns 1 if its condition is true and 0 otherwise. $p_{object|s_i}$ and $p_{human|s_i}$ are represented as follows:

$$p_{object|s_i} = \max(p_{that|s_i}, p_{which|s_i})$$
$$p_{human|s_i} = \max(p_{who|s_i}, p_{whose|s_i}, p_{whom|s_i})$$

Variables $p_{that|s_i}$, $p_{which|s_i}$, $p_{who|s_i}$, $p_{whose|s_i}$, and $p_{whom|s_i}$ are the probabilities of each relative pronoun substituting [MASK] in a given sentence. If the value of the first term in the Equation 3 is closer to 1, MLMs incorrectly predict higher probability of "which" or "that", and if the second term approaches 1, MLMs incorrectly predict higher probability of "who", "whose" or "whom". In other words, when the bias is close to 1, models tend to regard animals as objects; and if it is close to -1, they tend to treat them as humans.

To investigate the relationship between the bias represented in Equation 3 and the frequency bias in the corpora, we also calculate the correlation between the bias and the frequency of object-related pronouns ("that" and "which") referring to each animal name in Wikipedia and BookCorpus.

5. Experimental Results

In this section, we first report the results of the template-based experiments (Section 5.1), and then the results of the corpus-based experiment (Section 5.2).

5.1. Template-based Evaluation

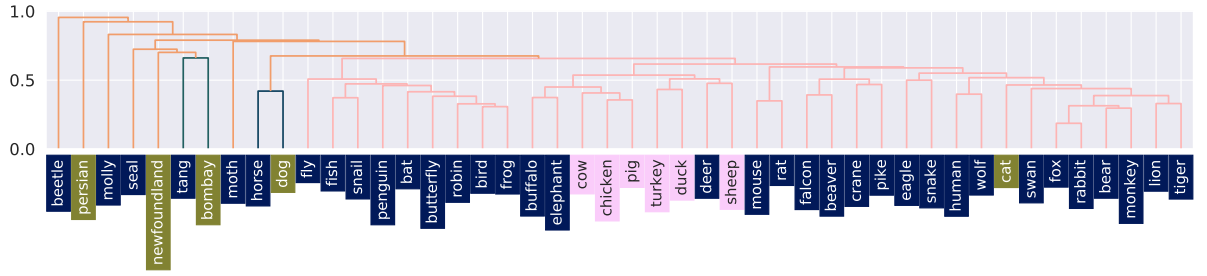
5.1.1. Probability Differences

The experimental results of probability differences between human and object sentences are presented in Figure 1, and Figures 4, 5, 7, 6 in Appendix A.

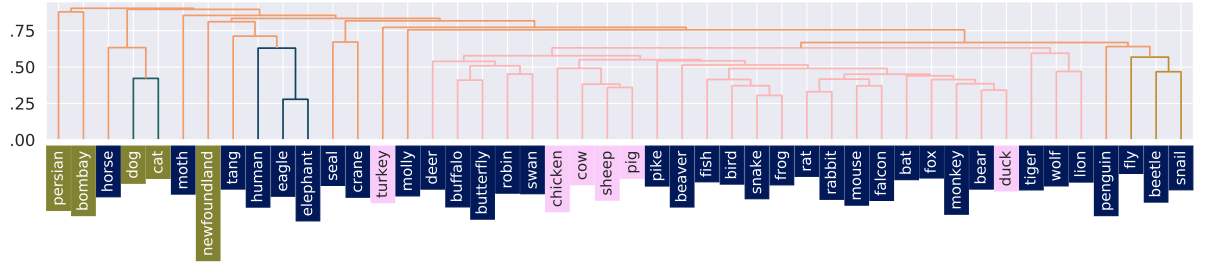
From Figure 1 it can be observed that the names of animals colored with the same color belong to roughly the same clusters. Especially in the results of BERT and RoBERTa, the names of animals who are often kept at farms were clustered closely in most cases (see Figures 1a and 1b). In the cases of DistilBERT and ALBERT, the animal names with the same color were not grouped together, but some belonged to the same cluster, indicating that they were not completely disjointed.

In Tables 2 and 3, we show sets of top five words with the largest probability change for each animal. For these tables we chose the five most frequent animal names in Wikipedia, and added the most popular animals living in farms and at homes, as they are one of the focal interest of our investigation: "cat", "dog", "chicken" and "pig". In these tables, we show the results for BERT and RoBERTa, while the results of the remaining models are given in Appendix A.

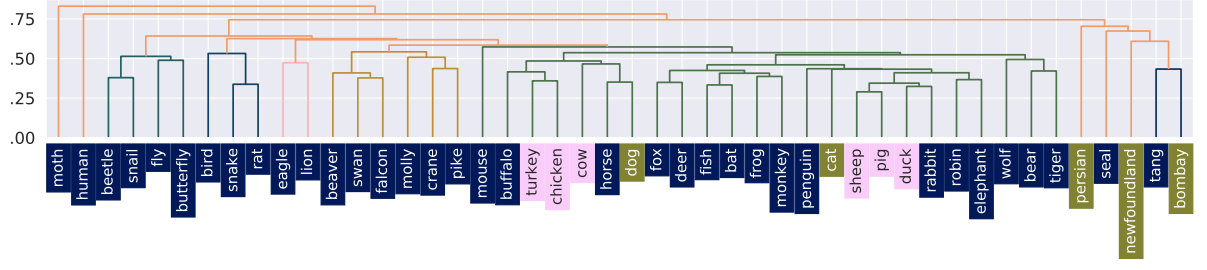
For "chicken", "pig" and "turkey", words with high probability change in object sentences included "slaughtered", "reproduced", "ripe" (see Table 2), also "dried" and "harvested" (see Table 3). Also, in BERT, "f**k"-rooted words were associated with many animals. On the other hand, in human sentences, associated words express personality and gender-related attributes, such as "clumsy" or "bisexual". There are also many words that represent personality traits that can be interpreted as negative, for example "sarcastic". On the other hand, "human" did not exhibit many such characteristics, so the bias observed in the case of nonhuman animals was not caused by a bias in the template sentence itself, but a bias among the sentence and the name.



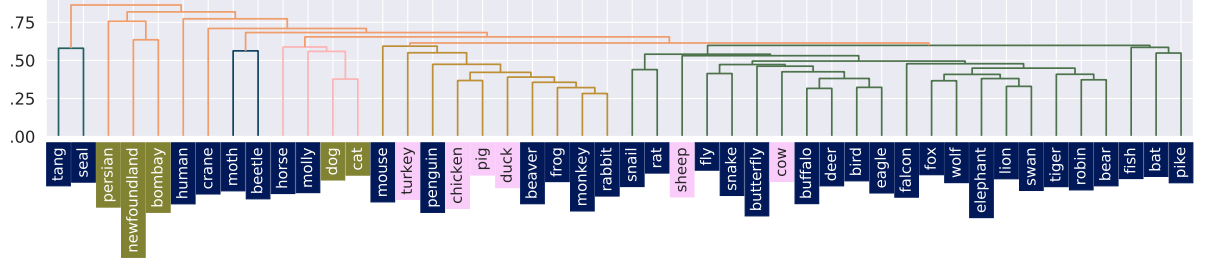
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) ALBERT

Figure 1: Results of hierarchical clustering based on the agreement rate of words whose predicted probability of filling the [MASK] token changed significantly between template sentences. The vertical axis represents 1−TMR by Equation 2. TMR indicates similarity among animal names. Each leaf is colored using SciPy library (Virtanen et al., 2020), with the default color threshold. ■ refers to "farm" animals, ■ indicates nonhuman companions and ■ stands for the remaining species.

Table 2
Sets of five predicted words with the highest change rate in BERT. Possibly harmful biased words are shown in **bold** font. For example, it is clear that "f**ked" and "slaughtered" can be harmful to non-human animals. Also, "ripe" and "beaf" can be harmful to some nonhuman animals, assuming that they are not food. refers to "farm" animals, indicates nonhuman companions and stands for the remaining species.

Animal name	Words with high probability change in *object* sentences	Words with high probability change in *human* sentences
cat	**f**ked**, **f**king**, **reproduced**, violated, **ripe**	**sarcastic**, **mute**, Ninja, **clumsy**, unnamed
dog	**f**ked**, **f**king**, struck, violated, committed	**sarcastic**, Ninja, **mute**, bisexual, unnamed
chicken	**slaughtered**, **f**ked**, **stamped**, **reproduced**, **ripe**	**clumsy**, **mute**, **sarcastic**, psychic, superhero
pig	**f**ked**, **stamped**, **slaughtered**, **reproduced**, **sin**	**clumsy**, **sarcastic**, **mute**, cheerful, blonde
turkey	**stamped**, **slaughtered**, **beef**, **ripe**, viable	**mute**, **clumsy**, psychic, **sarcastic**, **deaf**
fish	endemic, predatory, widespread, perennial, barred	heroine, **sarcastic**, Cinderella, princess, cheerful
fox	**f**ked**, happening, waking, calling, ours	**mute**, **sarcastic**, blonde, bisexual, **clumsy**
horse	**f**ked**, **sin**, violated, **stamped**, **ripe**	unnamed, pink, **sarcastic**, blonde, Ariel
human	ourselves, worth, ours, yours, our	bisexual, Ninja, **sarcastic**, blonde, lesbian

Table 3
Sets of five predicted words with the highest change rate in RoBERTa. Possibly harmful biased words are shown in **bold** font. refers to "farm" animals, indicates nonhuman companions and stands for the remaining species.

Animal name	Words with high probability change in *object* sentences	Words with high probability change in *human* sentences
cat	terrestrial, armoured, netted, scaled, predatory	foster, **deaf**, Transgender, Blind, Polish
dog	terrestrial, itself, predatory, defined, armoured	**deaf**, transsexual, foster, Homeless, lesbian
chicken	**dried**, freshwater, semen, **polled**, **harvested**	optimistic, **sarcastic**, romantic, pessimistic, Psychic
pig	**polled**, **dried**, **harvested**, yielded, **peeled**	romantic, **selfish**, optimistic, jealous, **arrogant**
turkey	**dried**, **processed**, ground, **slaughtered**, cached,	**deaf**, listening, jealous, optimistic, psychic
fish	freshwater, reef, widespread, **polled**, aggregate	**swearing**, jealous, witty, superhuman, sixteen
fox	**polled**, invasive, Madagascar, pictured, extant	pessimistic, **sarcastic**, **mercenary**, romantic, compassionate
horse	clicking, enough, beat, it, right	Transgender, lesbian, **deaf**, transgender, transsexual
human	extant, extinct, ours, yours, **edible**	bartender, nineteen, seventeen, sixteen, eighteen

Table 4
Frequency of relative pronouns referring to animal names in each corpus (references determined by CoreNLP). The number in parentheses is the total number minus the number indicating "human".

Corpus	that	which	who	whose	whom
Books3	104,244 (103,361)	28,552 (28,231)	44,607 (39,593)	4,115 (4,012)	2,006 (1,690)
Books-Corpus	5,111 (4,949)	1,470 (1,419)	3,925 (2,988)	183 (171)	66 (50)
Wikipedia (EN)	9,341 (9,265)	6,642 (6,586)	7,182 (6,648)	411 (396)	289 (274)

5.1.2. Sentiment Analysis

Next, we report the results of the sentiment analysis performed on obtained words in the experiment described in 4.1.2 (see Figure 2). The vertical axis of the figure shows the percentage of the number of words assigned to each sentiment. The horizontal one shows the sentiment and the names of the models.

We found that VADER assigned 0 (i.e. neutral sentiment) to the majority of the words, and that object sentences contained more neutral words than human sentences in all models. Contrary to our hypothesis, the ratio of negative words was found to be larger in human sentences for all three models except BERT. Within each model, the distribution of assigned sentiment was generally the same.

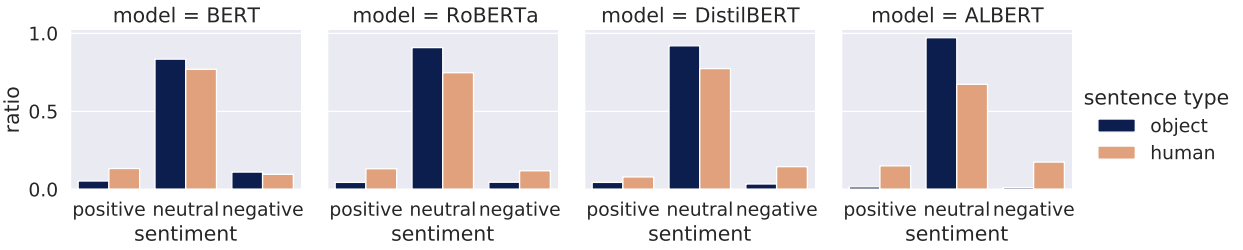
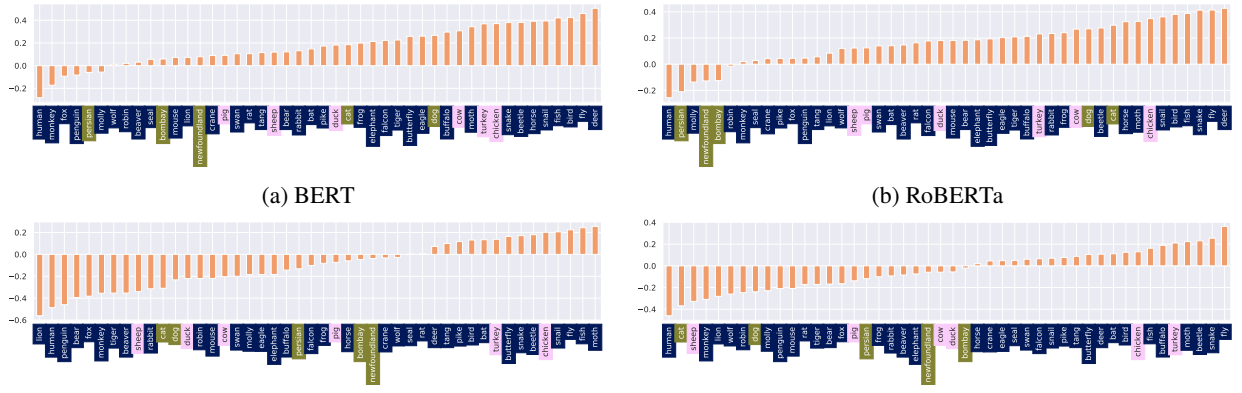


Figure 2: Results of sentiment analysis for each language model. Vertical axis shows the ratio of words assigned to a certain sentiment. For each sentiment, the darker bars indicate the percentage of words that have a higher mean probability in the object sentences, and the light-colored ones show the ratio of words that have a higher mean probability in the human sentences.



(a) BERT

(b) RoBERTa

(c) DistilBERT

(d) ALBERT

Figure 3: Results of the corpus-based bias analysis, sorted by the magnitude of the bias represented by Equation 3. Vertical axis shows the magnitude of the bias, where positive values indicate that MLMs incorrectly insert "that" or "which", and negative values indicate that MLMs incorrectly insert "who", "whose", or "whom" with higher probability. The horizontal one shows the animal names. Larger versions of the graphs are given in Appendix A. refers to "farm" animals, indicates nonhuman companions and stands for the remaining species.

5.2. Corpus-based Evaluation

Here, we present the results of the corpus-based experiment. First, we look at the sentences extracted from the corpora. In Table 4 we show the total numbers of relative pronouns referring to animal names in each corpus. The numbers in brackets indicate the total number minus the number of relative pronouns referring to "human". The total numbers for each animal are shown in Figures 8, 9 and 10. Comparing the total numbers of "that" and "which" with the total numbers of "who", "whose" and "whom", we found that the former group is about twice more common. This indicates that the corpus as a whole tends to treat nonhuman animals as objects. In addition, contrary to our assumption, the number of relative pronouns such as "who" that refers to "dogs" and "cats" in all corpora is almost the same as the total number of "that" and "which" (see Figures 8 and 9).

Next, we examine the results of analyzing the bias of MLMs using sentences collected from the Books3 corpus (see Figure 3). The vertical axis of each graph represents the degree of bias, and the horizontal one represents the animal names. A positive bias indicates a high probability of incorrectly entering "that" or "which" (i.e., having a speciesist bias), while a negative bias indicates a high probability of incorrectly filling "who", "whose", or "whom" (i.e., having a non-speciesist bias).

All of the models exhibited a negative bias against "human", and a positive bias against "chicken" and "turkey". These results are in line with our expectations. However, contrary to our predictions, the bias for "dog" and "cat" in BERT and RoBERTa is positive, indicating that they tend to be treated as objects. On the other hand, DistilBERT and ALBERT were found to include more negative bias, i.e. non-speciesist tendency, compared to BERT and RoBERTa. Table 5 shows the correlation between these biases and the ratio of the frequency of object-related pronouns in the

Table 5
Pearson correlation coefficient (r) between the bias represented in Equation 3 and frequency of object-related pronouns in Wikipedia and BookCorpus.

	BERT	RoBERTa	DistilBERT	ALBERT
r	0.77	0.55	0.81	0.74

Table 6
Pseudo-perplexity (PPPL) on pseudo-log-likelihood score (PLL) (Salazar et al., 2020) on 1,000 sentences, which are randomly sampled from collected sentences in Section 4.2. Lower value indicates better performance of mask predictions.

	BERT	RoBERTa	DistilBERT	ALBERT
PPPL	1.062	1.105	1.109	1.113

corpora. The correlation was above 0.7 for MLMs other than RoBERTa, and above 0.5 for RoBERTa, which indicates that the ratio of relative pronouns in the corpus explains the bias of MLMs to some extent. We think that the low value for RoBERTa is due to the fact that RoBERTa has been pre-trained on other corpora.

6. Discussion

6.1. Template-based Approach

The results of the animal names clustering in BERT and RoBERTa partially support the hypothesis of this experiment that MLMs vary animal-related words between object and human sentences depending on the animal name. On the other hand, DistilBERT and ALBERT performed clustering slightly different from our expectation, which may be due to the lower performance of mask predictions caused by the smaller model size.

To measure performance of mask predictions of each model, we calculate pseudo-perplexity (PPPL) on pseudo-log-likelihood score (PLL) (Salazar et al., 2020) of each model (see Table 6) on 1,000 sampled sentences from collected text for performing Corpus-based experiment (Section 4.2). BERT achieves the highest score, and other model's scores yield similar. These results may partially explain why BERT clustering mirrored our expectations. while RoBERTa clustering cannot be explained by the mask prediction performance.

As shown in Tables 2 and 3, when nonhuman animals are described by object sentences, they are linked with harmful words such as "f**ked". Furthermore, in the case of animals who live in farms to be utilized as flesh, meat-related words have been confirmed, for example "slaughtered" and "harvested" described as problematic in previous studies (Dunayer, 2001, 2003, see also Section 2.2). These words are likely to be associated with speciesist language that objectifies animals. In addition, contrary to the hypothesis of the present experiment, harmful words were also associated with human sentences. This suggests a speciesist bias in which MLMs treat nonhuman animals negatively, even if they are treated as humans.

In the experiments of sentiment analysis, it is important to note here that VADER itself may exhibit a speciesist bias. For example, VADER considers "killed" to be a negative word, but recognizes "slaughtered" as a neutral word. This problem should be investigated further.

6.2. Corpus-based Approach

Frequencies of human-related pronouns are lower than object-related pronouns in all corpora (see Table 4). There are at least two possible causes for this discrepancy: (1) there are fewer human-related relative pronouns that refer to nonhuman animals in the corpus than object-related ones, or (2) the recall of CoreNLP for human-related relative pronouns is low. If (1) is correct, it suggests that people tend to treat nonhuman animals as objects. If (2) is correct, it suggests that there is a bias in CoreNLP which makes the parser unable to sufficiently capture human-related relational references to nonhuman animals. Either result could be indirectly harmful to nonhuman animals.

In our corpus bias evaluation experiments, we found that, contrary to our hypothesis, the models had a speciesist bias against "dog" and "cat". However, all models exhibited a non-speciesist bias for more specific kinds of dogs and cats such as "newfoundland" and "persian". These results suggest that MLMs predicted "that" and "which" referring to "dog" and "cat" with high probability because they are commonly used as general names and therefore do not represent specific individuals. The bias between general names and more specific names will also be a subject of our future work.

6.3. Limitations

In all of these experiments, the speciesist bias is assessed using a difference between the object and human sentences. Therefore, it is not possible to evaluate biases that do not appear in that difference. Especially, we cannot evaluate words or concepts that are simply associated with animal names in this experiment. Since the purpose of this research is to investigate the relationship between animal names and speciesist language and its bias, it is necessary to evaluate other types of bias related to animal names in the future.

In addition, the corpus used in the corpus-based experiment is Books3, a corpus consisting of published books, so it is possible that different biases will be detected in different types of corpora. However, since the magnitude of this bias is correlated with the frequency of occurrence in Wikipedia and BookCorpus, we think that the same results are likely to be confirmed when using other textual resources.

7. How to Mitigate Speciesist Bias?

Based on the experimental results of this paper (and Hagendorff et al. (2022)), we briefly discuss possible ways to mitigate speciesist bias. Here we consider, according to the classification of the Shah et al. (2020), label bias, selection bias, overamplification and semantic bias as causes of the speciesist bias. Note that the ideal data distribution is a non-speciesist distribution, i.e. the strength of the relationship between nonhuman animals and negative concepts or speciesist language is significantly weaker than the strength of the relationship between positive concepts or non-speciesist language.

First, regarding the *label bias*, this bias emerges when there is discrepancy of the dependent variable (i.e. label) between source and the ideal distribution. Our experimental results are label-independent, so the speciesist bias found in this paper is not caused by the label bias. We did not analyze bias in downstream tasks such as sentiment analysis or text generation. However, it is possible to identify speciesism bias caused by label bias in the behavior of the NLP models used in downstream tasks. For example, as mentioned in Section 5.1.2, VADER (which is not a pre-trained language model) labels "slaughter" as neutral, meaning it contains a speciesist bias due to label bias. If a dataset for the sentiment analysis task contains a similar bias, then the behavior of models trained on that dataset may exhibit a speciesist bias. One way to mitigate the label bias is to create non-speciesist annotation guidelines for downstream tasks directly or indirectly related to nonhuman animals.

Second, regarding the *selection bias*, this bias occurs when there is a difference between source data distribution and ideal one. Current English resources are considered to be speciesist. Therefore, at least two mitigation methods are possible. The first method is to mitigate the bias of a corpus itself for example by using algorithms such as Counterfactual Data Augmentation (CDA) (Zhao et al., 2018; Webster et al., 2020). However, unlike in the cases of gender or race, because we do not know what attributes nonhuman animals correspond to, utilizing CDA would be problematic. The second approach is to deliberately select a corpus that uses non-speciesist language. Here, news articles and web content that ethical vegans tend to contribute may be useful. Devinney et al. (2020) has created Queer corpus from news and web content relating to LGBTQ+ people and their issues and found that the topics discussed in Queer corpus differ from the main stream English corpus. Perhaps, topics in textual data discussing ethical vegan-related topics also differ from the main stream English corpus, and there are fewer speciesist language.

Third, we consider *overamplification*, which occurs when, nevertheless the distribution of source data and ideal one are almost identical, a model amplifies small differences. The experimental results in this paper show that BERT, DistilBERT, and ALBERT associate different words with animal names, even though they were pre-trained on the same corpus. In particular, BERT associates more "f**k"-rooted words, and DistilBERT and ALBERT assign the same word to many animal names. This phenomenon is unlikely to be caused by word distribution in the corpus, and is probably mainly due to overamplification. On the other hand, RoBERTa appears to assign a wide variety of words, and since corpora other than Wikipedia and BookCorpus were used for pre-training RoBERTa, it may be important to pre-train language models on many diverse corpora.

Forth type of bias which could be mitigated is *semantic bias* which relates to "unintended or undesirable associations and social stereotypes" (Shah et al., 2020). Many methods for mitigating and analyzing this bias require specific word lists (e.g. Liang et al., 2020; Kaneko and Bollegala, 2021), but because some understanding of speciesism is needed to create them and many people are not even aware of the problem, recruting annotators would be difficult (also see the beginning of Section 4). Therefore, bias mitigation methods such as Dropout (Webster et al., 2020) and Self-debias (Schick et al., 2021), which do not require a specific word list, may be useful (cf. Meade et al., 2022).

Our experiments have also shown that biases differ among nonhuman animal species. Therefore, in addition to mitigating speciesist bias with humans, there is also a need to mitigate the bias among nonhuman animals. This may be difficult to do with methods described as above, therefore this problem should be addressed in the future.

8. Conclusion

In this paper, we analyzed the speciesist bias against animals inherent in MLMs. Our experimental results showed that such models strongly associate harmful words with many nonhuman animals. Especially for BERT and RoBERTa, these models clustered the animal names as we expected, and also associated the same harmful words (e.g. "slaughtered") with nonhuman animals who live in the farm. We also found that MLMs, especially BERT and RoBERTa, are biased to associate object-related pronouns ("that" and "which") with various nonhuman animals, and demonstrate that this bias is correlated with the frequency of these relative pronouns referring to each animal in the corpora.

Since this research is restricted to English language, it cannot be generalized to other languages. Moreover, this paper does not address so-called *intersectional bias*. For example, "bitch" means a female dog (intersectional attributes of gender and animal species), but it is also used as an insult toward women. In future, we plan to expand our research by utilizing findings in animal ethics regarding intersectional bias and discrimination between speciesist bias and other biases (Birke et al., 1995; Adams, 1990).

CRediT authorship contribution statement

Masashi Takeshita: Conceptualization of this study, Methodology, Software, Investigation, Data curation, Writing - Original Draft, Visualization. **Rafal Rzepka:** Writing - Review & Editing, Supervision. **Kenji Araki:** Supervision.

References

Adams, C.J., 1990. The sexual politics of meat: A feminist-vegetarian critical theory. Bloomsbury Publishing USA.

Bartl, M., Nissim, M., Gatt, A., 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain (Online). pp. 1–16. URL: https://www.aclweb.org/anthology/2020.gebnlp-1.1.

Basta, C., Costa-jussà, M.R., Casas, N., 2019. Evaluating the underlying gender bias in contextualized word embeddings, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy. pp. 33–39. URL: https://www.aclweb.org/anthology/W19-3805, doi:10.18653/v1/W19-3805.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 610–623. URL: https://doi.org/10.1145/3442188.3445922, doi:10.1145/3442188.3445922.

Birke, L., Dunayer, J., Kheel, M., 1995. Animals and women: Feminist theoretical explorations. Duke University Press.

Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H., 2020. Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 5454–5476. URL: https://www.aclweb.org/anthology/2020.acl-main.485, doi:10.18653/v1/2020.acl-main.485.

Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4356–4364.

Bordia, S., Bowman, S.R., 2019. Identifying and reducing gender bias in word-level language models, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Minneapolis, Minnesota. pp. 7–15. URL: https://aclanthology.org/N19-3002, doi:10.18653/v1/N19-3002.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 183–186. URL: https://science.sciencemag.org/content/356/6334/183, doi:10.1126/science.aal4230.

Crameri, F., 2021. Scientific colour maps. URL: https://doi.org/10.5281/zenodo.5501399, doi:10.5281/zenodo.5501399.

Crameri, F., Shephard, G.E., Heron, P.J., 2020. The misuse of colour in science communication. Nature communications 11, 1–10.

Dev, S., Li, T., Phillips, J.M., Srikumar, V., 2020. On measuring and mitigating biased inferences of word embeddings. Proceedings of the AAAI Conference on Artificial Intelligence 34, 7659–7666. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6267, doi:10.1609/aaai.v34i05.6267.

Dev, S., Li, T., Phillips, J.M., Srikumar, V., 2021a. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 5034–5050. URL: https://aclanthology.org/2021.emnlp-main.411, doi:10.18653/v1/2021.emnlp-main.411.

Dev, S., Sheng, E., Zhao, J., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Peng, N., Chang, K.W., 2021b. What do bias measures measure? arXiv preprint arXiv:2108.03362 .

Devinney, H., Björklund, J., Björklund, H., 2020. Semi-supervised topic modeling for gender bias discovery in English and Swedish, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain (Online). pp. 79–92. URL: https://aclanthology.org/2020.gebnlp-1.8.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423, doi:10.18653/v1/N19-1423.

Dunayer, J., 1995. Sexist words, speciesist roots, in: Animals and women: Feminist theoretical explorations. Duke University Press Durham, NC, pp. 11–31.

Dunayer, J., 2001. Animal Equality: Language and Liberation. Ryce Pub.

Dunayer, J., 2003. English and speciesism. English Today 19, 61–62. doi:10.1017/S0266078403001093.

Fischer, B., 2021. Animal Ethics: A Contemporary Introduction. New York: Routledge.

Franklin, E., 2020. Acts of killing, acts of meaning: an application of corpus pattern analysis to language of animal-killing. Ph.D. thesis. Lancaster University. URL: https://doi.org/10.17635/lancaster/thesis/931.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al., 2020. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 .

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J., 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences 115, E3635–E3644.

Garimella, A., Amarnath, A., Kumar, K., Yalla, A.P., N, A., Chhaya, N., Srinivasan, B.V., 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online. pp. 4534–4545. URL: https://aclanthology.org/2021.findings-acl.397, doi:10.18653/v1/2021.findings-acl.397.

Gokaslan, A., Cohen, V., 2019. Openwebtext corpus http://Skylion007.github.io/OpenWebTextCorpus.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., Lopez, A., 2021. Intrinsic bias metrics do not correlate with application bias, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 1926–1940. URL: https://aclanthology.org/2021.acl-long.150, doi:10.18653/v1/2021.acl-long.150.

Guo, W., Caliskan, A., 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. Computing Research Repository arXiv:2006.03955. URL: https://arxiv.org/abs/2006.03955.

Hagendorff, T., Bossert, L., Fai, T.Y., Singer, P., 2022. Speciesist bias in ai–how ai applications perpetuate discrimination and unfair outcomes against animals. arXiv preprint arXiv:2202.10848 .

Hellman, D., 2008. When is discrimination wrong? Harvard University Press.

Horta, O., 2018. Discrimination against vegans. Res Publica 24, 359–373.

Horta, O., Albersmeier, F., 2020. Defining speciesism. Philosophy Compass 15, e12708. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12708, doi:https://doi.org/10.1111/phc3.12708, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12708.

Hovy, D., Bianchi, F., Fornaciari, T., 2020. "you sound just like your father" commercial machine translation systems include stylistic biases, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 1686–1690. URL: https://aclanthology.org/2020.acl-main.154, doi:10.18653/v1/2020.acl-main.154.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S., 2020. Social biases in NLP models as barriers for persons with disabilities, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 5491–5501. URL: https://www.aclweb.org/anthology/2020.acl-main.487, doi:10.18653/v1/2020.acl-main.487.

Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media 8, 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Jepson, J., 2008. A linguistic analysis of discourse on the killing of nonhuman animals. Society & Animals 16, 127–148.

Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., Choi, Y., 2021. Delphi: Towards machine ethics and norms. arXiv preprint arXiv:2110.07574 .

Joseph, K., Morgan, J., 2020. When do word embeddings accurately reflect surveys on our beliefs about people?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 4392–4415. URL: https://www.aclweb.org/anthology/2020.acl-main.405, doi:10.18653/v1/2020.acl-main.405.

Kaneko, M., Bollegala, D., 2021. Debiasing pre-trained contextualised embeddings, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online. pp. 1256–1266. URL: https://aclanthology.org/2021.eacl-main.107, doi:10.18653/v1/2021.eacl-main.107.

Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y., 2019. Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy. pp. 166–172. URL: https://www.aclweb.org/anthology/W19-3823, doi:10.18653/v1/W19-3823.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2020. Albert: A lite bert for self-supervised learning of language representations, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=H1eA7AEtvS.

Lauscher, A., Glavaš, G., Ponzetto, S.P., Vulić, I., 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. Proceedings of the AAAI Conference on Artificial Intelligence 34, 8131–8138. URL: https://ojs.aaai.org/index.php/AAAI/article/

view/6325, doi:10.1609/aaai.v34i05.6325.

Lauscher, A., Lüken, T., Glavaš, G., 2021. Sustainable modular debiasing of language models. arXiv preprint arXiv:2109.03646 .

Li, T., Khashabi, D., Khot, T., Sabharwal, A., Srikumar, V., 2020. UNQOVERing stereotyping biases via underspecified questions, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online. pp. 3475–3489. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.311, doi:10.18653/v1/2020.findings-emnlp.311.

Liang, P.P., Li, I.M., Zheng, E., Lim, Y.C., Salakhutdinov, R., Morency, L.P., 2020. Towards debiasing sentence representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 5502–5515. URL: https://aclanthology.org/2020.acl-main.488, doi:10.18653/v1/2020.acl-main.488.

Liu, H., Dacon, J., Fan, W., Liu, H., Liu, Z., Tang, J., 2020. Does gender matter? towards fairness in dialogue systems, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online). pp. 4403–4416. URL: https://www.aclweb.org/anthology/2020.coling-main.390, doi:10.18653/v1/2020.coling-main.390.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized bert pretraining approach. Computing Research Repository arXiv:1907.11692. arXiv:1907.11692.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland. pp. 55–60. URL: https://www.aclweb.org/anthology/P14-5010, doi:10.3115/v1/P14-5010.

Manzini, T., Yao Chong, L., Black, A.W., Tsvetkov, Y., 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings, 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 615–621. URL: https://www.aclweb.org/anthology/N19-1062, doi:10.18653/v1/N19-1062.

May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R., 2019. On measuring social biases in sentence encoders, 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 622–628. URL: https://www.aclweb.org/anthology/N19-1063, doi:10.18653/v1/N19-1063.

Meade, N., Poole-Dayan, E., Reddy, S., 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland. pp. 1878–1898. URL: https://aclanthology.org/2022.acl-long.132, doi:10.18653/v1/2022.acl-long.132.

Menegatti, M., Rubini, M., 2017. Gender bias and sexism in language, in: Oxford Research Encyclopedia of Communication, Oxford University Press. URL: https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-470, doi:10.1093/acrefore/9780190228613.013.470.

Michener, C.D., Sokal, R.R., 1957. A quantitative approach to a problem in classification. Evolution 11, 130–162.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Curran Associates Inc., Red Hook, NY, USA. p. 3111–3119.

Nadeem, M., Bethke, A., Reddy, S., 2021. StereoSet: Measuring stereotypical bias in pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416, doi:10.18653/v1/2021.acl-long.416.

Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R., 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 1953–1967. URL: https://www.aclweb.org/anthology/2020.emnlp-main.154, doi:10.18653/v1/2020.emnlp-main.154.

Owe, A., Baum, S.D., 2021. Moral consideration of nonhumans in the ethics of artificial intelligence. AI and Ethics , 1–12.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation, in: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Presser, S., 2020. Books3 https://twitter.com/theshawwn/status/1320282149329784833.

Regan, T., 2004. The case for animal rights. Univ of California Press.

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., Kalai, A., 2019. What's in a name? Reducing bias in bios without access to protected attributes, 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4187–4195. URL: https://www.aclweb.org/anthology/N19-1424, doi:10.18653/v1/N19-1424.

Salazar, J., Liang, D., Nguyen, T.Q., Kirchhoff, K., 2020. Masked language model scoring, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 2699–2712. URL: https://aclanthology.org/2020.acl-main.240, doi:10.18653/v1/2020.acl-main.240.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 .

Schick, T., Udupa, S., Schütze, H., 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. Transactions of the Association for Computational Linguistics 9, 1408–1424. URL: https://doi.org/10.1162/tacl_a_00434, doi:10.1162/tacl_a_00434, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00434/1979270/tacl_a_00434.pdf.

Sealey, A., Pak, C., 2018. First catch your corpus: methodological challenges in constructing a thematic corpus. Corpora 13, 229–254. URL: https://doi.org/10.3366/cor.2018.0145, doi:10.3366/cor.2018.0145, arXiv:https://doi.org/10.3366/cor.2018.0145.

Shah, D.S., Schwartz, H.A., Hovy, D., 2020. Predictive biases in natural language processing models: A conceptual framework and overview, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 5248–5264. URL: https://aclanthology.org/2020.acl-main.468, doi:10.18653/v1/2020.acl-main.468.

Sheng, E., Chang, K.W., Natarajan, P., Peng, N., 2019. The woman worked as a babysitter: On biases in language generation, 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), Association for Computational Linguistics, Hong Kong, China. pp. 3407–3412. URL: https://www.aclweb.org/anthology/D19-1339, doi:10.18653/v1/D19-1339.

Sheng, E., Chang, K.W., Natarajan, P., Peng, N., 2021. Societal biases in language generation: Progress and challenges, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 4275–4293. URL: https://aclanthology.org/2021.acl-long.330, doi:10.18653/v1/2021.acl-long.330.

Silva, A., Tambwekar, P., Gombolay, M., 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 2383–2389. URL: https://aclanthology.org/2021.naacl-main.189, doi:10.18653/v1/2021.naacl-main.189.

Singer, P., 2015. Animal Liberation. Vintage Digital.

Stanczak, K., Augenstein, I., 2021. A survey on gender bias in natural language processing. arXiv preprint arXiv:2112.14168 .

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, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy. pp. 1630–1640. URL: https://www.aclweb.org/anthology/P19-1159, doi:10.18653/v1/P19-1159.

Takeshita, M., Rafal, R., Araki, K., 2021. [Analysis of speciesist bias inherent in the english language model] Eigo no gengo moderu ni naizaisuru shusabetsu baiasu no bunseki (in Japanese), in: Proceedings of The 27th Annual Meeting of The Association for Natural Language Processing (NLP-2021), Association for Natural Language Processing. pp. 1–6. URL: https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/A1-1.pdf.

Tan, S., Joty, S., Kan, M.Y., Socher, R., 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 2920–2935. URL: https://aclanthology.org/2020.acl-main.263, doi:10.18653/v1/2020.acl-main.263.

Tan, Y.C., Celis, L.E., 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations, in: Wallach, H., Larochelle, H., Beygelzimer, A., d\textquotesingle Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 13230–13241. URL: https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf.

Trinh, T.H., Le, Q.V., 2018. A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Petrov, S., 2020. Measuring and reducing gendered correlations in pre-trained models. Computing Research Repository arXiv:2010.06032. arXiv:2010.06032.

Yoon, C.K., 2009. Naming nature: the clash between instinct and science. WW Norton & Company.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W., 2019. Gender bias in contextualized word embeddings, 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 629–634. URL: https://www.aclweb.org/anthology/N19-1064, doi:10.18653/v1/N19-1064.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W., 2018. Gender bias in coreference resolution: Evaluation and debiasing methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana. pp. 15–20. URL: https://www.aclweb.org/anthology/N18-2003, doi:10.18653/v1/N18-2003.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, USA. p. 19–27. URL: https://doi.org/10.1109/ICCV.2015.11.

A. Appendix

Table 7
Hyperparameter information regarding models using all experiments in this paper

	$BERT_{LARGE\text{-}cased}$	$RoBERTa_{LARGE}$	$DistilBERT_{base\text{-}cased}$	$ALBERT_{large\text{-}v2}$
Parameters	334M	334M	65M	18M
Layers	24	24	6	24
Hidden	1024	1024	768	1024
Embedding	1024	1024	768	128
Other	-	-	distillation model of BERT	using parameter sharing

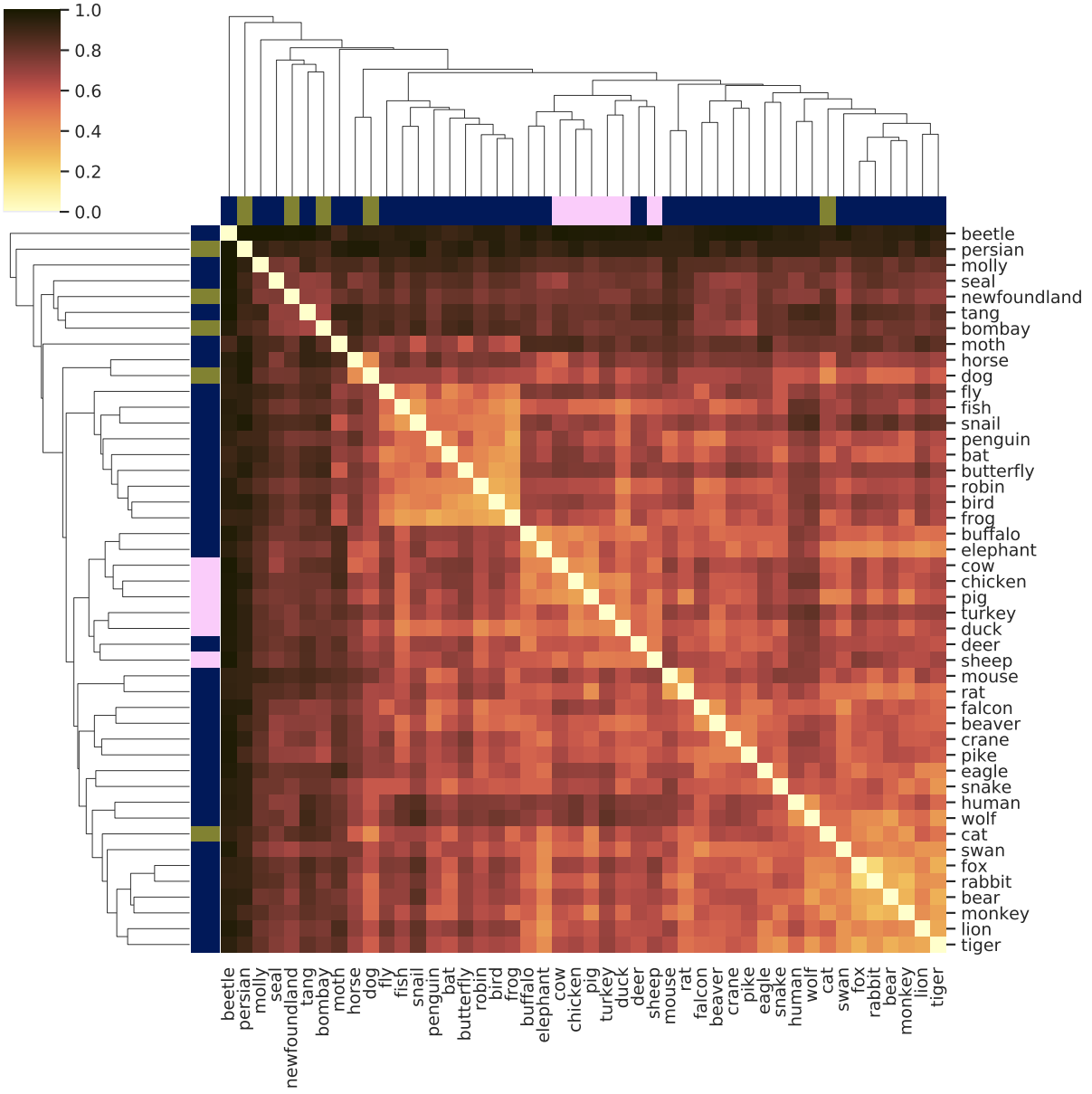


Figure 4: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in BERT: refers to "farm" animals, indicates nonhuman companions and stands for the remaining species. The score represents the 1−TMR. The lighter cells, the higher similarity among animals.

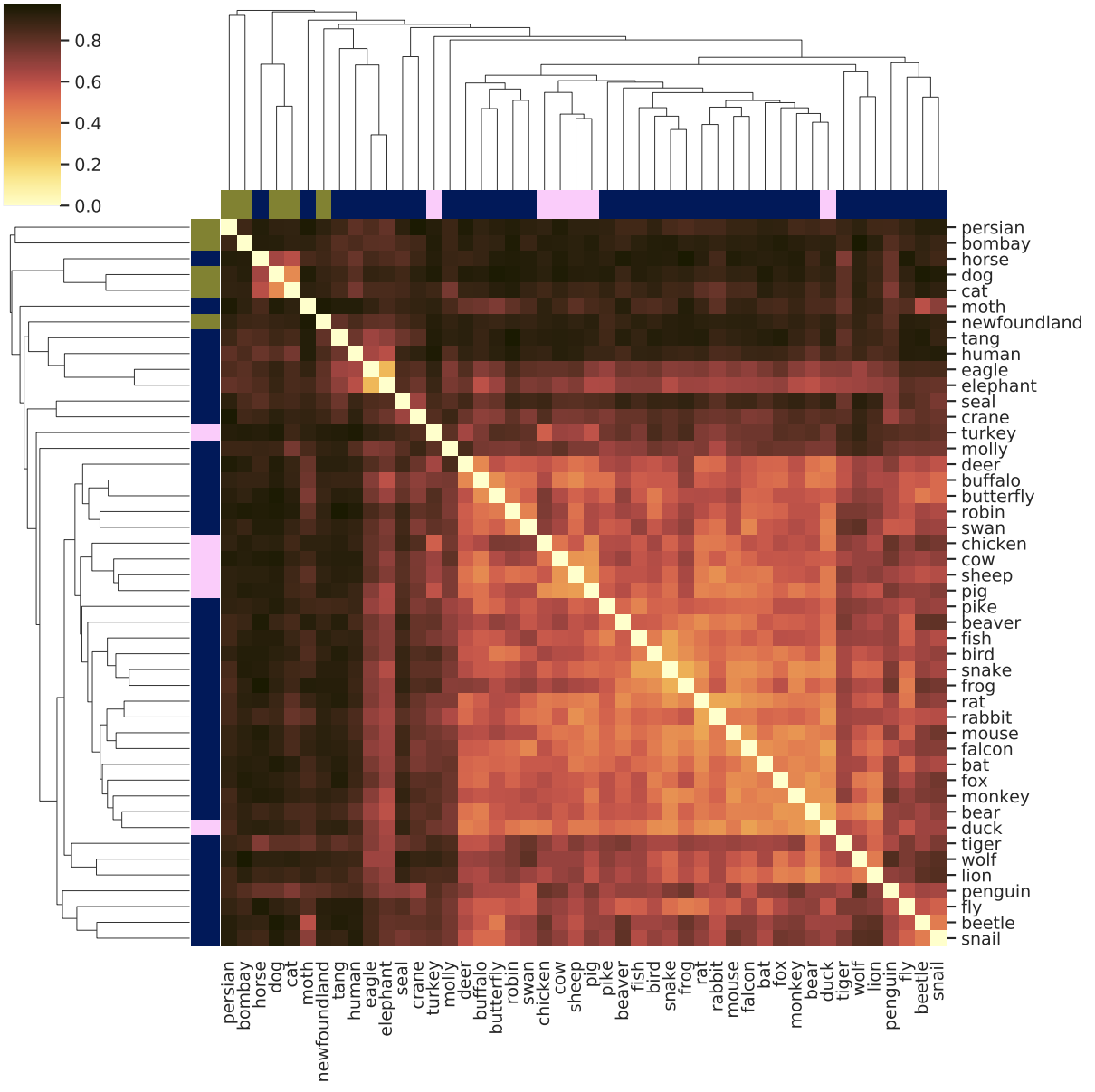


Figure 5: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in RoBERTa: refers to "farm" animals, indicates nonhuman companions and stands for the remaining species. The score represents the 1−TMR. The lighter cells, the higher similarity among animals.

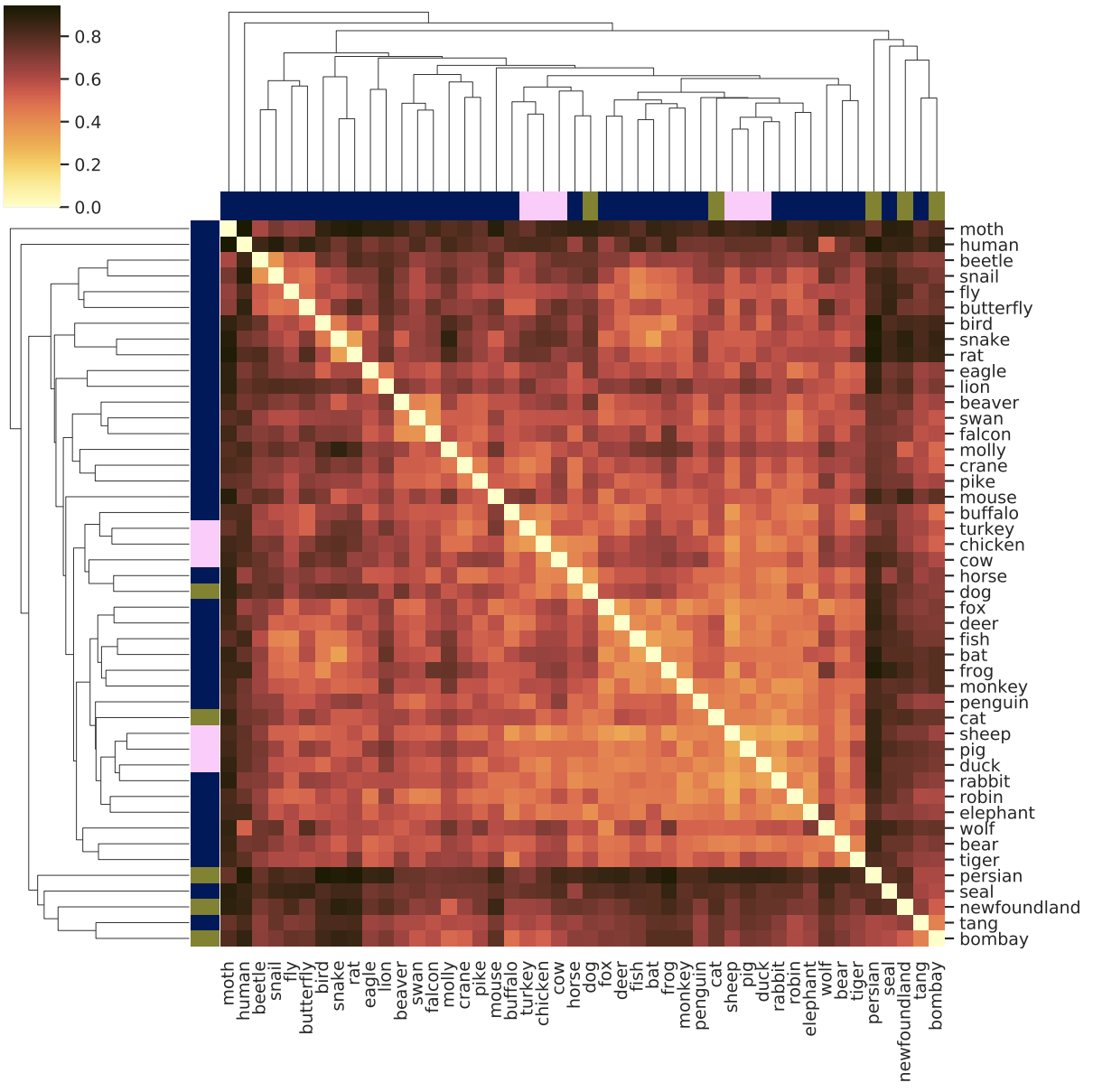


Figure 6: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in DistilBERT: refers to "farm" animals, indicates nonhuman companions and stands for the remaining species. The score represents the 1−TMR. The lighter cells, the higher similarity among animals.

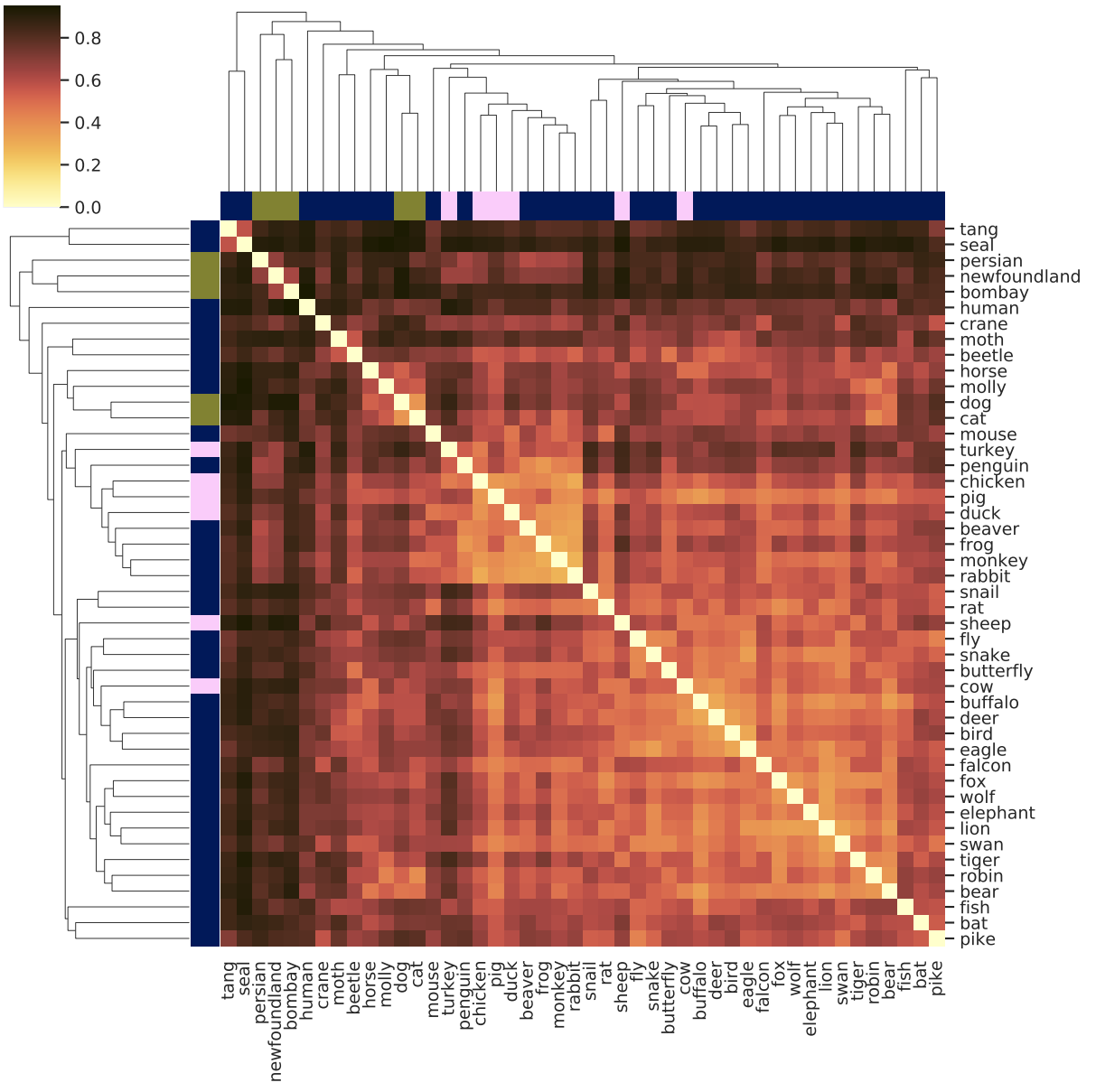


Figure 7: A heat map of the results of the template-based experiments, clustered by TMR with large probability changes in ALBERT: refers to "farm" animals, indicates nonhuman companions and stands for the remaining species. The score represents the 1−TMR. The lighter cells, the higher similarity among animals.

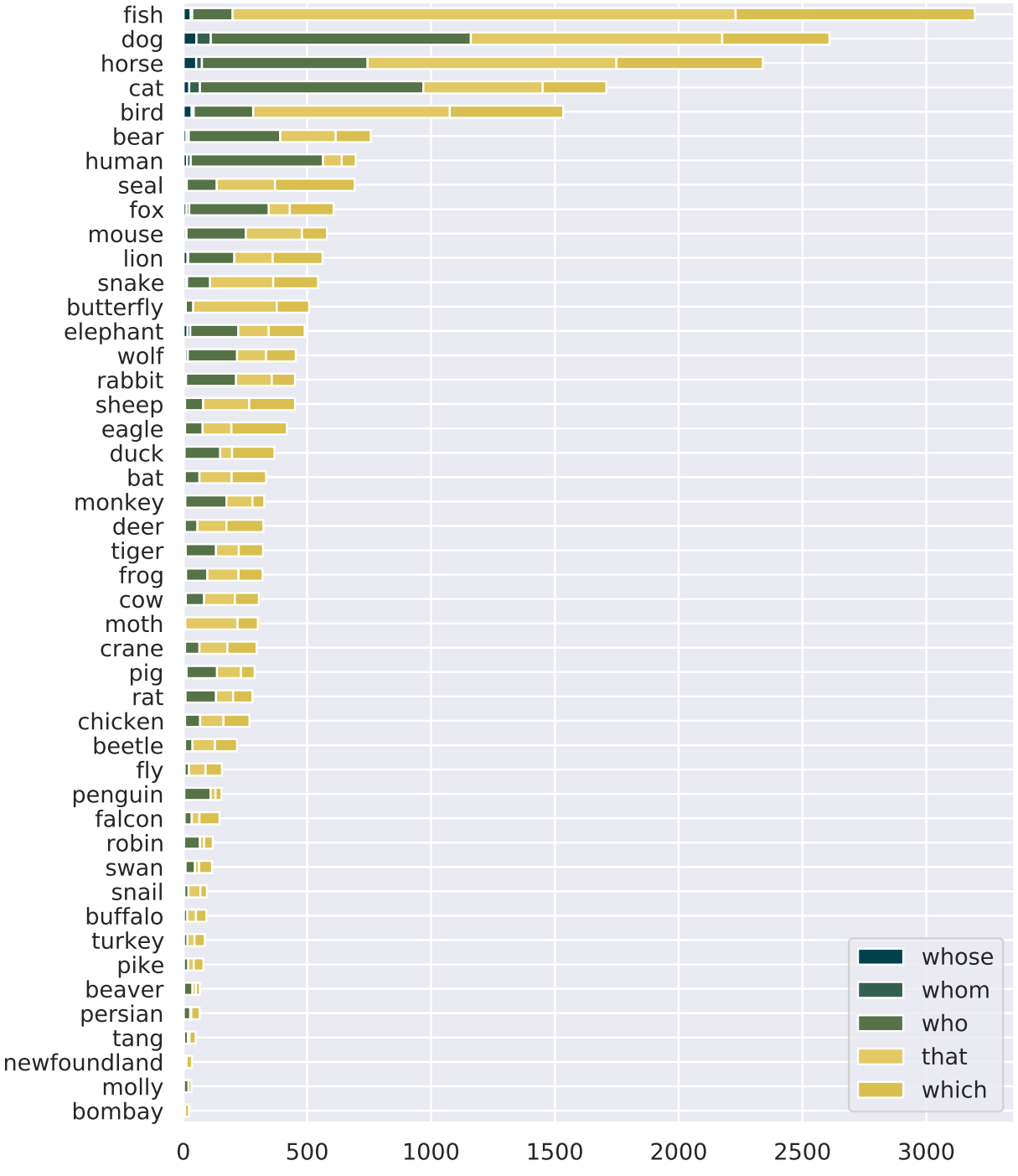


Figure 8: Number of relative pronouns referring to each animal in English Wikipedia

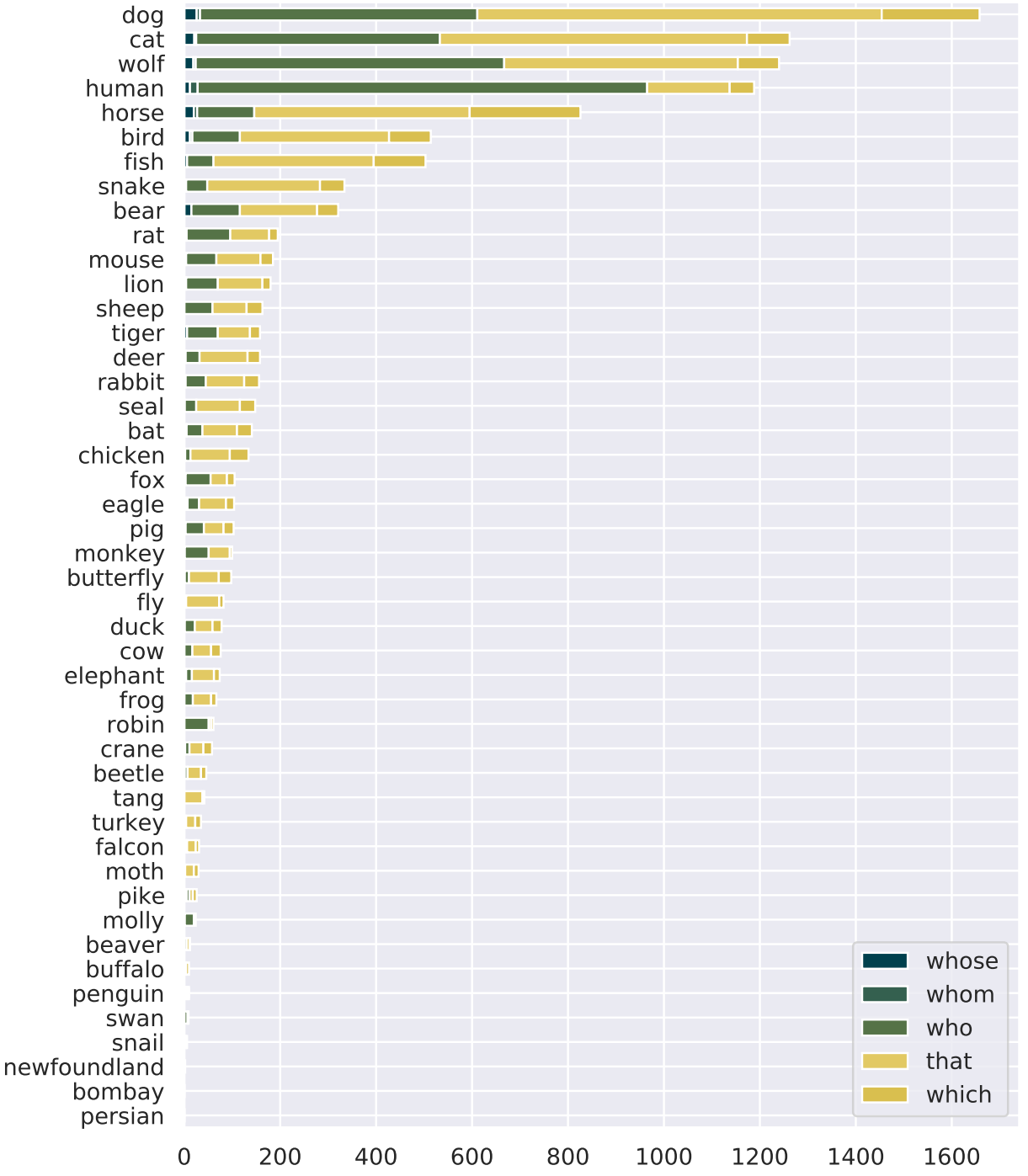


Figure 9: Number of relative pronouns referring to each animal in BookCorpus

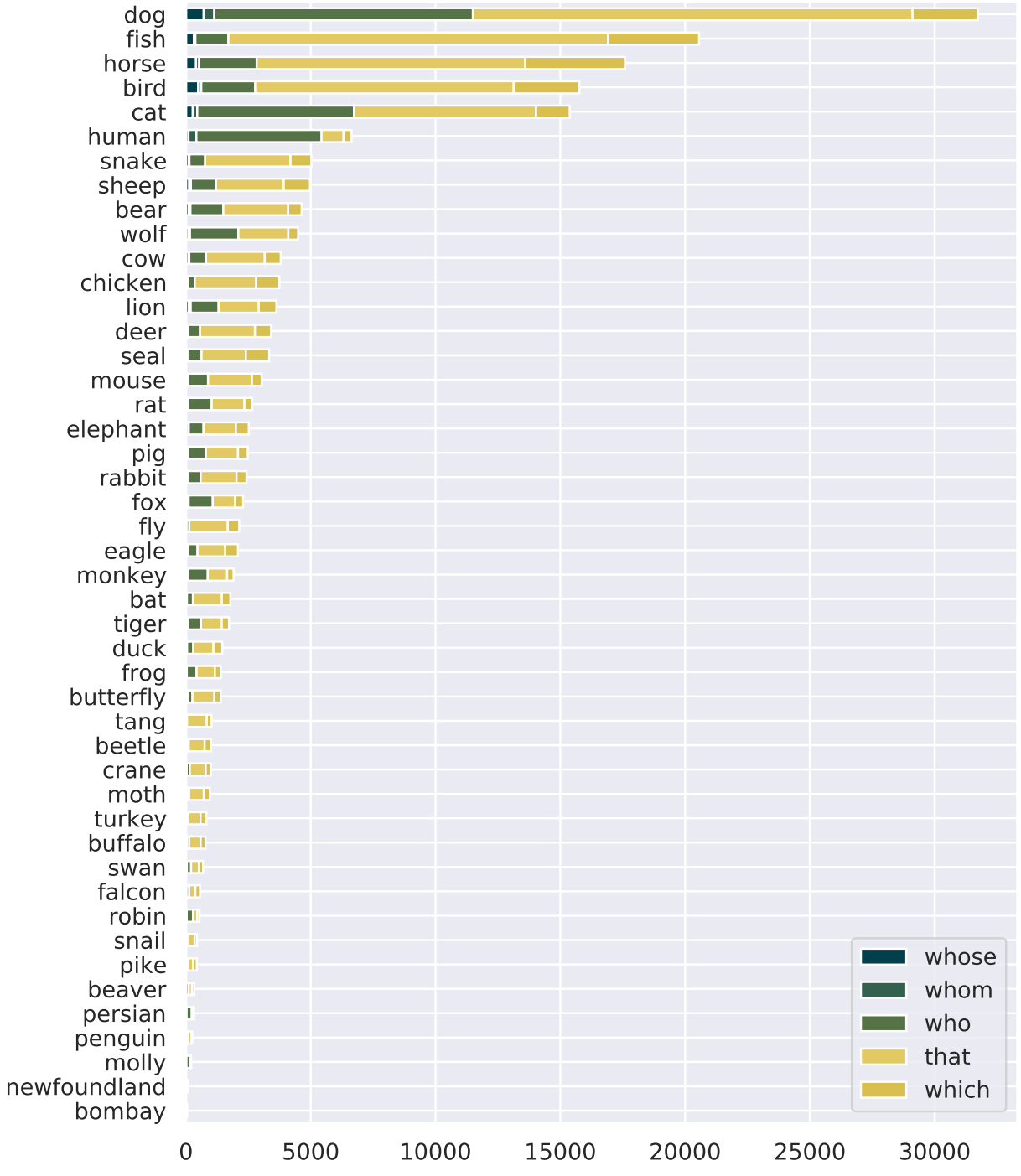


Figure 10: Number of relative pronouns referring to each animal in Books3

Table 8
Sets of five words with the highest change rate in BERT

animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	endemic, threatened, predatory, barred, endangered	Ninja, sarcastic, blonde, Nordic, psychic
bear	f**ked, ours, waking, happening, stirring	sarcastic, bisexual, mute, psychic, blonde
beaver	endemic, reproduced, f**ked, extinct, viable	mute, psychic, Ninja, sarcastic, superhero
beetle	conspicuous, stinging, waking, ripe, variable	coach, coaching, coaches, Swiss, midfielder
bird	endemic, threatened, uncommon, endangered, widespread	sarcastic, blonde, psychic, superhero, heroine
bombay	standardized, portable, timed, ceremonial, audible	unemployed, widowed, homeless, heroine, psychologist
buffalo	stamped, f**ked, beef, slaughtered, reproduced	psychic, mute, sarcastic, clumsy, blind
butterfly	endemic, widespread, uncommon, threatened, disputed	sarcastic, superhero, blonde, cheerful, mute
cat	f**ked, f**king, reproduced, violated, ripe	sarcastic, mute, Ninja, clumsy, unnamed
chicken	slaughtered, f**ked, stamped, reproduced, ripe	clumsy, mute, sarcastic, psychic, superhero
cow	f**ked, stamped, slaughter, slaughtered, ripe	sarcastic, mute, clumsy, psychic, cheerful
crane	loading, rotating, operating, overhead, tuned	psychic, blonde, sarcastic, mute, heroine
deer	beef, f**ked, f**king, viable, barred	mute, fairy, sarcastic, Ariel, psychic
dog	f**ked, f**king, struck, violated, committed	sarcastic, Ninja, mute, bisexual, unnamed
duck	endemic, f**ked, reproduced, endangered, viable	sarcastic, psychic, clumsy, mute, cheerful
eagle	barred, circling, happening, endemic, yours	mute, psychic, sarcastic, Amazon, blonde
elephant	f**ked, stamped, reproduced, f**king, happening	sarcastic, mute, psychic, clumsy, cheerful
falcon	barred, endemic, f**ked, reproduced, extinct	blonde, Ninja, psychic, sarcastic, Amazon
fish	endemic, predatory, widespread, perennial, barred	heroine, sarcastic, Cinderella, princess, cheerful
fly	predatory, toxic, stinging, endemic, colonial	sarcastic, cheerful, superhero, blonde, genius
fox	f**ked, happening, waking, calling, ours	mute, sarcastic, blonde, bisexual, clumsy
frog	endemic, threatened, endangered, ##olate, widespread	sarcastic, Ninja, cheerful, clumsy, blonde
horse	f**ked, sin, violated, stamped, ripe	unnamed, pink, sarcastic, blonde, Ariel
human	ourselves, worth, ours, yours, our	bisexual, Ninja, sarcastic, blonde, lesbian
lion	ours, happening, waking, arising, pictured	psychic, sarcastic, mute, heroine, bisexual
molly	unacceptable, theirs, treason, happening, occurring	blonde, mute, deaf, cheerful, widowed
monkey	f**ked, f**king, waking, happening, ours	sarcastic, mute, clumsy, lesbian, Ninja
moth	endemic, Crambidae, ##tropical, variable, Geometridae	unemployed, Ninja, DJ, psychic, undefeated
mouse	reproduced, viable, mating, f**ked, endemic	sarcastic, cheerful, clumsy, Dorothy, superhero
newfoundland	ours, theirs, happening, paradise, nearer	bisexual, protagonist, narrator, heroine, blonde
penguin	endemic, extinct, endangered, barred, reproduced	sarcastic, psychic, Ninja, clumsy, mute
persian	periodic, convex, contraction, symmetric, bounded	deaf, genius, widowed, ##headed, intelligent
pig	f**ked, stamped, slaughtered, reproduced, sin	clumsy, sarcastic, mute, cheerful, blonde
pike	endemic, barred, preferred, edged, subspecies	blonde, mute, widowed, cheerful, homeless
rabbit	f**ked, waking, happening, slaughtered, arriving	mute, sarcastic, bisexual, psychic, clumsy
rat	reproduced, f**ked, viable, reared, waking	sarcastic, mute, clumsy, Gothic, cheerful
robin	endemic, subspecies, threatened, barred, unmistakable	mute, sarcastic, psychic, cheerful, mechanic
seal	stamped, forged, valid, void, binding	Brave, blonde, Ninja, psychic, mute
sheep	endemic, f**ked, sustainable, perennial, viable	mute, sarcastic, psychic, princess, narrator
snail	predatory, endemic, widespread, fossil, marine	sarcastic, cheerful, mute, optimistic, psychic
snake	endemic, yours, barred, ours, venom	sarcastic, cheerful, blonde, mute, optimistic
swan	yours, ours, f**ked, reproduced, endemic	psychic, sarcastic, mute, mechanic, clumsy
tang	audible, repeated, nasal, consonant, pronounced	Smart, unemployed, smart, homeless, brave
tiger	happening, ours, f**king, waking, f**ked	mute, sarcastic, psychic, bisexual, blonde
turkey	stamped, slaughtered, beef, ripe, viable	mute, clumsy, psychic, sarcastic, deaf
wolf	ours, yours, happening, waking, you	bisexual, mute, sarcastic, psychic, lesbian

Table 9
Sets of five words with the highest change rate in RoBERTa

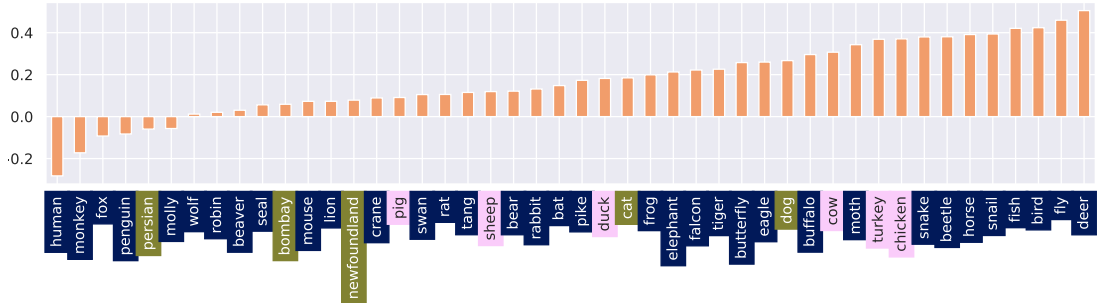
animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	intact, handled, dried, unloaded, batted	virtuous, heroic, witty, superhuman, princess
bear	polled, extant, freshwater, handled, endemic	superhuman, mercenary, romantic, sarcastic, prince
beaver	invasive, freshwater, widespread, dried, common	atheist, lonely, swearing, nineteen, lesbian
beetle	deposited, feeding, circulating, hardest, clustered	virtuous, heroic, fictional, philosophical, courageous
bird	freshwater, offshore, migr, endemic, extant	Human, philosophical, jealous, sarcastic, witty
bombay	fallacy, phosphorus, absurdity, gelatin, FALSE	Shy, loyal, married, shy, wealthy
buffalo	dried, freshwater, listed, polled, stamped	cowardly, arrogant, selfish, cunning, rebellious
butterfly	variable, common, offshore, widespread, clustered	virtuous, superhuman, philosophical, rebellious, heroic
cat	terrestrial, armoured, netted, scaled, predatory	foster, deaf, Transgender, Blind, Polish
chicken	dried, freshwater, semen, polled, harvested	optimistic, sarcastic, romantic, pessimistic, Psychic
cow	polled, dried, semen, domestically, processed	romantic, optimistic, witty, poetic, mysterious
crane	erected, automated, propelled, loader, towed	jealous, psychic, horny, deaf, conflicted
deer	bucks, dried, harvested, roadside, buck	swearing, witty, romantic, philosophical, jealous
dog	terrestrial, itself, predatory, defined, armoured	deaf, transsexual, foster, Homeless, lesbian
duck	freshwater, polled, dried, offshore, netted	superhuman, heroic, superhero, protagonist, Human
eagle	correlated, achievable, warranted, measurable, irreversible	adventurer, hacker, Paladin, Sailor, trainer
elephant	achievable, warranted, happening, extinct, irreversible	adventurer, detective, Lesbian, thief, vigilante
falcon	freshwater, netted, largest, aerial, perched	Human, optimistic, superhuman, rebellious, lesbian
fish	freshwater, reef, widespread, polled, aggregate	swearing, jealous, witty, superhuman, sixteen
fly	respiratory, common, genital, dried, larvae	heroic, lonely, witty, Talking, intuitive
fox	polled, invasive, Madagascar, pictured, extant	pessimistic, sarcastic, mercenary, romantic, compassionate
frog	freshwater, larvae, widespread, invasive, dart	superhuman, seventeen, nineteen, swearing, heroic
horse	clicking, enough, beat, it, right	Transgender, lesbian, deaf, transgender, transsexual
human	extant, extinct, ours, yours, edible	bartender, nineteen, seventeen, sixteen, eighteen
lion	pictured, Madagascar, Guinea, polled, Bengal	Human, prince, princess, mercenary, Princess
molly	edible, larvae, harvested, dried, invasive	pessimistic, Persian, nineteen, deaf, lazy
monkey	polled, palm, Madagascar, extant, Guinea	virtuous, mercenary, superhuman, Alone, romantic
moth	happening, circulating, newer, collapsing, getting	prophetic, divine, :, Blind, feminist
mouse	polled, larvae, extant, freshwater, edible	swearing, romantic, Alone, heroic, rich
newfoundland	Antarctica, unfolding, contiguous, ours, wetlands	deaf, transsexual, bisexual, runner, addicted
penguin	lower, offshore, flattened, freshwater, oval	lesbian, unmarried, married, rebellious, feminist
persian	larvae, edible, peeled, citrus, vegetation	atheist, writer, novelist, journalist, physicist
pig	polled, dried, harvested, yielded, peeled	romantic, selfish, optimistic, jealous, arrogant
pike	freshwater, offshore, invasive, Atlantic, harvested	Human, protector, nineteen, optimistic, swearing
rabbit	dried, widespread, netted, terrestrial, harvested	sarcastic, Psychic, optimistic, pessimistic, heroic
rat	dried, freshwater, widespread, polled, extant	heroic, swearing, superhuman, romantic, protector
robin	common, variable, migrating, widespread, larvae	superhuman, virtuous, philosophical, trustworthy, irresponsible
seal	tightening, tightened, tighter, stamped, dried	autistic, Hungry, dreaming, deaf, transsexual
sheep	polled, dried, harvested, yielded, processed	jealous, witty, arrogant, heroic, optimistic
snail	minute, deposited, dried, flattened, occurring	clueless, Psychic, jealous, cowardly, loyal
snake	freshwater, netted, dried, invasive, widespread	superhuman, swearing, cursed, immortal, protagonist
swan	freshwater, aerial, lower, largest, netted	protector, trustworthy, forgiving, pessimistic, loyal
tang	contraction, residue, correlation, causation, correlated	deaf, homeless, transsexual, Homeless, veterinarian
tiger	manageable, corrected, viable, right, largest	lesbian, princess, transsexual, vegan, Human
turkey	dried, processed, ground, slaughtered, cached	deaf, listening, jealous, optimistic, psychic
wolf	polled, extant, heaviest, widespread, invasive	Psychic, wizard, Human, Loki, prince

Table 10
Sets of five words with the highest change rate in DistilBERT

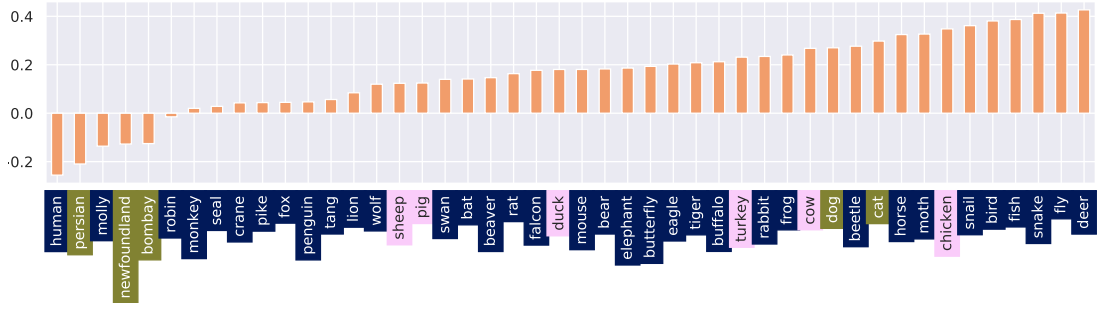
animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	endemic, distributed, widespread, ##olate, ##gratory	magician, psychic, witch, villains, wizard
bear	endemic, distributed, valid, edible, convex	psychic, witches, witch, herself, grandmother
beaver	distributed, endemic, lateral, ##gratory, inactivated	psychic, heroine, archaeologist, magician, narrator
beetle	endemic, widespread, distributed, subsp, valid	magician, transgender, psychic, widowed, deaf
bird	endemic, distributed, widespread, variable, declining	robot, psychic, princess, witches, angel
bombay	quarterly, annual, administered, recited, yearly	widowed, transgender, deaf, bisexual, blind
buffalo	endemic, abolished, extinct, inactivated, edible	heroine, actress, girlfriend, psychic, narrator
butterfly	endemic, widespread, distributed, valid, decreasing	lion, psychic, controlling, gifted, vain
cat	endemic, valid, convex, inactivated, viable	narrator, thirteen, psychic, fourteen, seventeen
chicken	endemic, edible, pounded, differentiated, clarified	deaf, blind, psychic, narrator, bullying
cow	endemic, edible, differentiated, sacred, branched	homeless, deaf, bullying, blind, paranoid
crane	distributed, towed, valid, endemic, unfolded	psychic, heroine, deaf, magician, actress
deer	endemic, ##gratory, distributed, extinct, subspecies	psychic, sailor, narrator, witch, grandmother
dog	endemic, subspecies, branched, differentiated, valid	herself, teenage, widowed, thirteen, grandmother
duck	endemic, valid, distributed, subspecies, edible	psychic, narrator, clumsy, deaf, thirteen
eagle	endemic, distributed, valid, lateral, decreasing	psychic, princess, witches, herself, fairies
elephant	endemic, distributed, convex, valid, inhabited	heroine, magician, nurse, psychic, princess
falcon	convex, scaled, lateral, distributed, endemic	psychic, transgender, magician, controlling, kidnapped
fish	endemic, distributed, widespread, variable, diagnostic	widowed, narrator, sailor, genius, girlfriend
fly	distributed, valid, extant, endemic, occurring	deaf, motorcycle, sailor, narrator, thirteen
fox	endemic, distributed, ##gratory, extinct, extant	heroine, magician, psychic, sailor, narrator
frog	endemic, distributed, widespread, variable, valid	vain, princess, fairies, psychic, narrator
horse	valid, equivalent, propelled, endemic, assessed	heroine, grandmother, witches, fairies, princess
human	worth, acceptable, our, reproduced, valid	princess, witch, emerald, angel, witches
lion	endemic, engraved, displayed, seated, valid	psychic, heroine, princess, witches, witch
molly	frequented, underway, inhabited, unfinished, excavated	bisexual, deaf, transgender, elderly, widowed
monkey	endemic, distributed, valid, convex, differentiated	princess, witch, magician, herself, witches
moth	widespread, occurring, varies, irregular, subsp	blind, sighted, deaf, blinded, astronomer
mouse	distributed, endemic, inactivated, valid, bilateral	witches, fairies, thirteen, prostitutes, witch
newfoundland	endemic, populated, inhabited, frequented, dotted	widowed, secretary, bisexual, transgender, pregnant
penguin	endemic, valid, distributed, extinct, extant	psychic, widowed, narrator, magician, actress
persian	convex, bounded, periodic, continuous, compact	transgender, actress, widowed, nurse, wrestler
pig	endemic, edible, viable, differentiated, inactivated	thirteen, dolls, narrator, girlfriend, seventeen
pike	valid, longitudinal, endemic, distributed, convex	psychic, heroine, fairies, caring, narrator
rabbit	endemic, distributed, viable, differentiated, inactivated	magician, psychic, witch, witches, narrator
rat	endemic, distributed, oral, lateral, bilateral	witches, witch, fairies, wizard, princess
robin	endemic, distributed, valid, branched, widespread	psychic, heroine, autism, deaf, narrator
seal	stamped, filed, valid, worn, engraved	heroine, psychic, kidnapped, protagonist, drowning
sheep	endemic, distributed, inactivated, viable, extinct	psychic, housekeeper, narrator, thirteen, witch
snail	widespread, distributed, endemic, variable, minute	psychic, villain, widowed, protagonist, lion
snake	distributed, endemic, ##olate, variable, diagnostic	witch, princess, fairies, wizard, goddess
swan	endemic, distributed, ##tail, lateral, ##gratory	psychic, narrator, magician, transgender, autism
tang	recited, oral, cumulative, meaningful, elastic	blind, widowed, heroine, deaf, scientist
tiger	endemic, distributed, valid, extant, inhabited	heroine, widowed, princess, lovers, witch
turkey	endemic, extant, edible, widespread, valid	deaf, actress, psychic, transgender, narrator
wolf	endemic, valid, conspicuous, edible, variable	witches, witch, princess, fairies, grandmother

Table 11
Sets of five words with the highest change rate in ALBERT

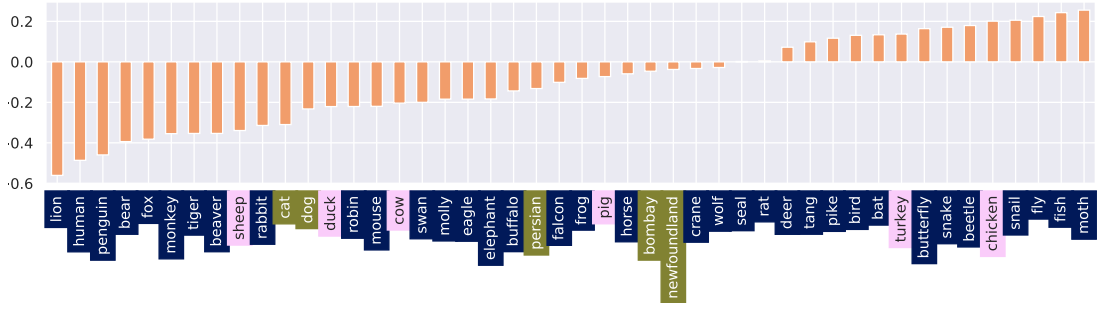
animal name	words with high probability change in *object* sentences	words with high probability change in *human* sentences
bat	printed, basalt, lodged, cylindrical, mandible	confident, gambler, dreamer, fearless, grieving
bear	lodged, reported, indicated, suggested, excavated	trusting, helpless, fearless, trusted, obedient
beaver	noticeable, lodged, brownish, coughed, yellowish	heiress, dreamer, princess, bachelor, addict
beetle	leaked, brownish, lodged, occurring, yellowish	princess, adventurer, dreamer, valkyrie, knighted
bird	printed, brownish, yellowish, lodged, localized	dreamer, conqueror, princess, angels, slaves
bombay	reopened, commenced, redeveloped, expanded, skyline	widow, soprano, eunuch, knighted, pregnant
buffalo	brownish, lodged, yellowish, reported, basalt	dreamer, hero, helpless, obedient, widow
butterfly	printed, yellowish, brownish, highlighted, forewing	dreamer, conqueror, adventurer, himself, superhuman
cat	lodged, boar, appeared, urine, yellowish	dreamer, fearless, bachelor, confident, jed
chicken	spelt, brownish, lodged, compressed, stemmed	dreamer, atheist, jealous, telepathic, princess
cow	weigh, spelt, raked, brownish, lodged	destiny, conqueror, happiness, trusting, fearless
crane	corrugated, aluminium, hangar, diameter, turbine	jealous, dreamer, eunuch, homosexual, tigre
deer	brownish, lodged, reported, yellowish, surfaced	dreamer, slaves, conqueror, trusting, angels
dog	lodged, boar, suggested, reported, spelt	perfection, caring, fearless, loving, faithful
duck	spelt, contains, termed, spelled, containing	atheist, adventurer, dreamer, addict, estranged
eagle	resembled, printed, brachy, tapered, holotype	conqueror, helpless, widow, steward, dreamer
elephant	reported, brownish, yellowish, lodged, surfaced	helpless, conqueror, obedient, slaves, estranged
falcon	resembled, compressed, mandible, resembles, rectangular	dreamer, fearless, trusting, addict, obedient
fish	tapered, formulated, brownish, stemmed, tasted	dreamer, jealous, himself, atheist, conqueror
fly	nitrogen, printed, tapered, compressed, brownish	conqueror, dreamer, hostage, slaves, murderer
fox	brownish, yellowish, dorsal, bluish, puma	dreamer, trusted, slaves, trusting, selfish
frog	resembled, contains, spelt, termed, compressed	atheist, princess, bachelor, transgender, dreamer
horse	hoof, suggested, raked, overturned, hydraulic	caring, fearless, helpless, trusting, perfection
human	http, suggested, computed, spelt, stated	savior, loves, protector, loving, beloved
lion	noticeable, indicated, reported, yellowish, conical	trusting, estranged, helpless, dreamer, selfish
molly	spelled, suggested, advertised, yellowish, bacterio	confidant, dreamer, confident, obedient, fearless
monkey	resembled, spelt, xylo, termed, suggested	estranged, princess, dreamer, atheist, bachelor
moth	widespread, annual, biennial, localized, basal	dreamer, jed, magician, sorcerer, himself
mouse	generate, termed, contains, kernel, xml	wealthy, fearless, princess, billionaire, dreamer
newfoundland	happen, happened, place, reopened, resumed	widow, knighted, pregnant, transgender, addict
penguin	contained, brownish, noticeable, smelled, yellowish	widow, atheist, addict, billionaire, heiress
persian	quartz, sodium, clarified, indicated, contrary	heiress, wealthy, married, widow, unmarried
pig	brownish, termed, dorsal, yellowish, spelt	trusting, jealous, princess, selfish, estranged
pike	corrugated, diameter, tapered, aluminium, compressed	jealous, gambler, grieving, helpless, dreamer
rabbit	snout, resembled, contains, termed, spelt	dreamer, atheist, princess, transgender, estranged
rat	nitrogen, termed, contains, 1:, containing	dreamer, selfish, conqueror, estranged, atheist
robin	plumage, brownish, yellowish, printed, spelt	dreamer, confident, heroine, psychopath, selfish
seal	minimize, compress, tissue, membrane, corrugated	temeraire, racehorse, knighted, valkyrie, shepherd
sheep	aerobic, discontinued, uploaded, dorsal, reported	traitor, helpless, trusting, conqueror, slaves
snail	nitrogen, termed, corrugated, sodium, containing	selfish, dreamer, strangers, helpless, obedient
snake	localized, bluish, yellowish, pointed, brownish	messiah, conqueror, dreamer, helpless, obedient
swan	printed, tapered, erupted, conical, plumage	helpless, obedient, trusting, conqueror, estranged
tang	nitrogen, minimize, termed, compressed, compress	adventurer, conqueror, abbess, empress, barbarian
tiger	yellowish, brownish, bluish, excavated, reported	helpless, trusting, selfish, caretaker, incapable
turkey	tasted, dried, sliced, crisp, highlighted	adventurer, atheist, transgender, telepathic, knighted
wolf	mandible, dorsal, conical, termed, brownish	dreamer, helpless, princess, orphan, traitor



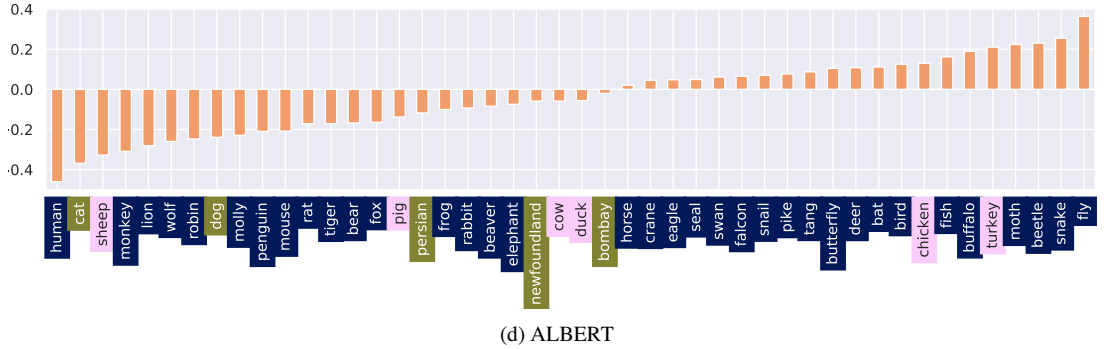
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) ALBERT

Figure 11: Results of the corpus-based bias analysis, sorted by the magnitude of the bias represented by Equation 3. Vertical axis shows the magnitude of the bias, where positive values indicate that MLMs incorrectly insert "that" or "which", and negative values indicate that MLMs incorrectly insert "who", "whose", or "whom" with higher probability. The horizontal one shows the animal names.