Multilingual and Multimodal Bias Probing and Mitigation in Natural Language Processing

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

eingereicht von
Victor Steinborn

München, den 14. November 2023

Erstgutachter: *Prof. Dr. Hinrich Schütze*

Zweitgutachter: *Prof. Dr. Aram Galstyan*

Tag der Einreichung: 14. November 2023

Tag der mündlichen Prüfung: 15. April 2024

Eidesstattliche Versicherung
(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5.)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig ohne unerlaubte Beihilfe angefertigt ist.

München, den 14. November 2023

Victor Steinborn

Abstract

Gender bias is a key global challenge of our time according to the United Nations sustainability goals, which call for the elimination of all forms of gender-based discrimination. Since it is ubiquitous online and offline, gender bias is also prevalent in the training data for Natural Language Processing models; these models therefore learn and internalize this bias. Gender bias then reappears when models are probed and used in downstream tasks such as automatic recruitment leading to gender-based discrimination that affects people's lives in a negative way. Thus, gender bias is problematic as it harms individuals.

There is a growing body of research attempting to combat gender bias in language models. However, the diversity of research is quite limited and focused on English and on occupational biases. In this thesis, we attempt to move beyond the current insular state of gender bias research in language models to improve the coverage of languages and biases that are being studied.

Specifically, we undertake three projects that aim to broaden the breadth of current gender bias research in Natural Language Processing (NLP). The first project aims to build a dataset to investigate languages beyond English; our methodology makes it easy to extend the dataset to any language of choice. In addition, we propose a new analytical bias measure that may be used to evaluate bias, given the model's prediction probabilities. In the second project, we demonstrate that learned gender stereotypes regarding politeness may bleed into cyberbullying detection systems, which may disproportionately fail to protect women if the system is attacked with honorifics. In this project, we focus on Korean and Japanese NLP models; however, our results raise the question whether other systems in other languages can fall prey to the same biases. In the third project, we demonstrate that visual representations of emoji may evoke harmful text generation that disproportionately affects different genders, depending on the emoji choice.

Zusammenfassung

Geschlechtsspezifische Vorurteile sind laut den Nachhaltigkeitszielen der Vereinten Nationen, die die Beseitigung aller Formen der geschlechtsspezifischen Diskriminierung fordern, eine der wichtigsten globalen Herausforderungen unserer Zeit. Da sie online und offline allgegenwärtig sind, sind geschlechtsspezifische Vorurteile auch in den Trainingsdaten für Modelle der natürlichen Sprachverarbeitung weit verbreitet. Diese Modelle lernen und verinnerlichen daher diese Vorurteile. Die geschlechtsspezifischen Vorurteile kommen dann wieder zum Vorschein, wenn die Modelle geprüft und in nachgelagerten Aufgaben wie der automatischen Rekrutierung verwendet werden, was zu geschlechtsspezifischer Diskriminierung führt, die das Leben der Menschen negativ beeinflusst. Daher sind geschlechtsspezifische Vorurteile problematisch, da sie dem Einzelnen schaden.

Es gibt eine wachsende Zahl von Forschungsarbeiten, die versuchen geschlechtsspezifische Vorurteile in Sprachmodellen zu bekämpfen. Allerdings ist die Vielfalt der Forschung recht begrenzt und konzentriert sich auf Englisch und auf berufliche Vorurteile. In dieser Arbeit versuchen wir, die derzeitige Insellage der Forschung zu geschlechtsspezifischen Vorurteilen in Sprachmodellen zu überwinden und die Abdeckung der untersuchten Sprachen und Vorurteile, die untersucht werden, auf eine breitere Basis zu stellen.

Konkret führen wir drei Projekte durch, die darauf abzielen, die Breite der aktuellen Gender-Bias-Forschung im Bereich Natural Language Processing (NLP) zu erweitern. Das erste Projekt zielt darauf ab, einen Datensatz zur Untersuchung von Sprachen außerhalb des Englischen zu erstellen. Unsere Methodik macht es einfach, den Datensatz auf jede beliebige Sprache zu erweitern. Darüber hinaus schlagen wir ein neues analytisches Bias-Maß vor, das zur Bewertung des Bias angesichts der Vorhersagewahrscheinlichkeiten des Modells verwendet werden kann. Im zweiten Projekt zeigen wir, dass erlernte Geschlechterstereotypen in Bezug auf Höflichkeit in Cybermobbing-Erkennungssysteme einfließen können, die Frauen möglicherweise unverhältnismäßig nicht schützen, wenn das System mit Ehrentiteln angegriffen wird. In diesem Projekt konzentrieren wir uns auf koreanische und japanische NLP-Modelle; unsere Ergebnisse werfen jedoch die Frage auf, ob auch andere Systeme in anderen Sprachen denselben Vorurteilen zum Opfer fallen können. Im dritten Projekt zeigen wir, dass visuelle Darstellungen von Emojis eine schädliche Textgenerierung hervorrufen können, die je nach Emoji-Auswahl unterschiedliche Geschlechter unverhältnismäßig stark betrifft.

Acknowledgments

This thesis is dedicated to my family, my girlfriend, friends and kind teachers who supported me throughout my life and the course of my PhD. I am immensely grateful for their unwavering support and this thesis would not be possible without them.

Publications and Declaration of Co-Authorship

Chapter 2 corresponds to the following publication:

> **Victor Steinborn**, Philipp Dufter, Haris Jabbar, and Hinrich Schütze. 2022. An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

I conceived of the original research contributions under the guidance of my advisor Hinrich Schütze. I conducted all practical and theoretical work on the paper with the following exceptions: Philipp Dufter performed the initial literature review, Haris Jabbar recruited annotators and the annotators annotated the published dataset. I wrote the initial draft of the article and did most of the subsequent corrections. I regularly discussed this work with my co-authors who assisted me in improving the manuscript.

Chapter 3 corresponds to the following preprint:

> **Victor Steinborn**, Antonis Maronikolakis, and Hinrich Schütze. 2023. Politeness Stereotypes and Attack Vectors: Gender Stereotypes in Japanese and Korean Language Models. arXiv:2306.09752.

I conceived of the original research contributions and performed all implementations and evaluations except for the conception and implementation of the training methodology for the debiasing method (Antonis Maronikolakis). I wrote the initial draft of the article and regularly discussed this work with my co-authors who assisted me in improving the paper.

Chapter 4 corresponds to an unpublished project with the following authors:

Victor Steinborn and Hinrich Schütze.

I conceived of the original contributions and performed all research. I wrote the initial draft of the chapter. My advisor Hinrich Schütze, assisted me in improving the chapter.

Contents

Chapter 1

Introduction

> Dreams are the mind's version of reality perfected.
>
> *NieR: Automata*
> *Jean-Paul*

In this thesis we attempt to tackle gender bias in modern Natural Language Processing (NLP) from novel perspectives. We propose a multilingual method to investigate gender bias across languages, that may be expanded to any language of interest. In addition, we also propose a bias measure based on information theory that lends itself to simple interpretation, designed to avoid statistical problems associated with small dataset sizes, which are common in social bias studies. Furthermore, we demonstrate modern NLP systems are susceptible to gender bias on the basis of politeness, using Korean and Japanese, and that the performance of cyber bullying detection systems are also susceptible to such biases. We additionally propose a simple solution to mitigating the harm in such cyber bullying detection systems, by simply providing politeness attacks to the models as training data. Finally, we investigate how textual and visual representations of emoji may impact text generation models and demonstrate that these models are susceptible to generating harmful text, which disproportionately affects one gender over the other, depending on the emoji.

1.1 Motivation

Gender equality is one of the world's key global challenges, as outlined by the United Nations' (UN) sustainability goals and the initiatives centered around it (United Nations General Assembly, 2015). Traditionally, gender equality focuses on the educational opportunities, economic empowerment, political representation, employment opportunities and general empowerment of women and girls (United Nations General Assembly, 2015). Especially important is the elimination of all forms of discrimination against women, with this being listed as the first goal under the gender equality sustainability goal (United Nations General Assembly, 2015). The engagement of men and boys, to meet this goal is also highlighted (United Nations General Assembly, 2015). In this work we focus on all forms of bias on the basis of gender, which we will refer to as gender bias in this work.

While the importance of gender equality may be apparent in society, the relevance of gender bias in NLP systems may not be readily apparent. Modern NLP systems are trained on large corpora from the internet, which contain historical social biases, which may bleed into NLP models (Caliskan et al., 2017; Sun et al., 2019). In particular, the prevalence of gender bias on the internet is substantial; for example, on Twitter, posts that exhibit gender bias are most common out of the posts that exhibit a form of social bias (Sap et al., 2020). These learned biases in turn may affect the performance of an NLP system in a downstream task (Blodgett et al., 2020). One high-stakes setting where the propagation of these biases is undesirable is in recruitment (Sun et al., 2019; De-Arteaga et al., 2019). In automated recruiting, an NLP model searches through potential candidates' professional profiles to determine if they would be a good fit for the role that is being hired for (De-Arteaga et al., 2019). A well-known example of an experimental automated recruiting system is that of Amazon, as reported by Reuters, which was shown to assign more negative scores to resumes that contained the word "women's", in the sense of "women's soccer team" (Jeffrey Dastin, 2018). While this particular system was never implemented in production, it serves to highlight the dangers and allocational harms of real-life algorithmic discrimination.

Other than allocational harms, which cover the systematic biased distribution of resources (such as employment opportunities in the previous example), representational harms cover biases concerning how people are represented by NLP models (Blodgett et al., 2020). For example, coreference resolution systems have been shown to perform poorly when resolving female pronouns with male-dominated professions, as it reinforces the stereotype that women don't participate in male-dominated jobs (Rudinger et al., 2018). Representational biases grounded in society are harmful in their own right as NLP models reproduce and amplify these social biases, effectively acting as a warped social mirror (Blodgett et al., 2020; Jia et al., 2020).

In connection with representational biases, representation in a more broad sense is another important factor to consider when studying gender bias. Gender equality is a global challenge, as evidenced by it being a UN sustainability goal, as outlined earlier (United Nations General Assembly, 2015). However, in the training data of large modern NLP models, while vast in size and diversity, especially when using large internet corpora like the Common Crawl[1], certain demographics are overrepresented while others are underrepresented (Bender et al., 2021). In particular, young men from developed countries are the largest contributor to the training data, while women and people from developing countries contribute less (Pew Research Center, 2021; The World Bank, 2021; Michael, 2020; Bender et al., 2021). For example, a recent survey of Wikipedia contributors revealed that only 8-15% of Wikipedia authors are female (Michael, 2020). Thus, large NLP models trained on such datasets may suffer from conforming to dominant hegemonic viewpoints, limiting the representation of the views of women and other demographics in such models (Bender et al., 2021).

[1] https://commoncrawl.org/

Generalizing the discussion, expanding on the issues relating to training data representation, more general issues with training AI models themselves make the task of training responsible AI models challenging. Although more focused on the behavior of reinforcement learning agents, the field of AI alignment arguably has significant overlap with bias research as it deals with building models that are consistent with human values (such as ethics) (Hendrycks et al., 2022; Gabriel, 2020). One key challenge in AI alignment is specification gaming (otherwise known as proxy gaming), where models simply optimize an objective function at the expense of not conforming to the desired human values of the users or developers of the model (Hendrycks et al., 2022; Gabriel, 2020). For example, chatbots like ChatGPT (OpenAI, 2022) are trained to imitate text from its training set while also maximizing the approval of human evaluators and safety systems. However, despite performing well on these systems, the early implementation of ChatGPT was well known to generate factually incorrect text that can even fool experts and to propagate undesirable gender biases, despite the presence of these safety systems (i.e. the safety systems were "gamed", the model did not actually learn the desired values) (Ji et al., 2023; Gao et al., 2023; Davey, 2022). Hence, eliminating gender bias in NLP systems may also be viewed as an alignment problem, i.e. aligning the learning objectives of the model with the desired output, that is consistent with human values.

While the topic of bias research is vast, we attempt to tackle it from various novel angles in the context of NLP. Before we outline our contributions, we will outline the limitations of current gender bias research that inspired our research projects, outlined in this thesis.

1.2 Current Limitations of Gender Bias Research

In this section we break the current limitations of bias research down into categories. These are areas we identified as current research gaps we aimed to address in our research. We will go through each category in turn and briefly discuss how we contributed to alleviating these issues in our projects.

1.2.1 English-Centric Research

One of the most prevalent limitations of current bias research is its English-centrism. Most research papers on gender bias are written to analyze English text with some works focusing on non-English languages; however, these are relatively rare and focus on specific (mostly European) languages (González et al., 2020; Bartl et al., 2020; Liang et al., 2020; Nozza et al., 2021).

This English-centrism is not necessarily a problem for some languages, as there are several techniques that lend themselves to being generalized. For example Liang et al. (2020) made use of English templates that can be directly translated into Chinese to test for gender biases in masked language models. Similarly, Nozza et al. (2021) utilized several European languages to investigate hurtful sentence completion in autoregressive models, such as GPT-2 (Radford et al., 2018b).

However, while there are some successes in directly transferring methods that work in English, more often there are situations where directly transferring the English technique to other languages is problematic or is simply impossible without a significant alteration of the technique. For example, in the work of Bartl et al. (2020), which attempts to make use of templates to measure gender bias in masked language models, English templates were directly translated to German, however due to gender agreement rules in German, the templates needed to be modified depending on the gender of the person being referred to in the sentence. It is suspected that due to this complication, their proposed technique is ineffective for German and yields results that do not reflect real-world biases (Bartl et al., 2020). Thus, in general, techniques may not be simply translated from English to German, and by extension other languages with similar gender agreement rules, as more sophisticated techniques are required.

Finally, we would also like to note that even works that are extremely critical of current trends in bias research, such as that of Blodgett et al. (2020), fail to acknowledge the English centrism of current research. However, given that the mitigation of gender bias is a global challenge, as evidenced by its inclusion in the UN sustainability goals, we see the inclusion of other languages as a vital step to this end (United Nations General Assembly, 2015). We will attempt to tackle this issue by building a multilingual gender bias challenge dataset, and by studying gender bias on the basis of politeness using Japanese and Korean in our projects.

1.2.2 Limited Coverage of Biases

Gender bias in NLP is often studied in connection with occupations. Most studies investigate the relative prediction probabilities of gendered person terms (such as “he” or “she”) in reference to persons described by occupations, such as “programmer” or “secretary” (Bartl et al., 2020; Bhaskaran and Bhallamudi, 2019; Liang et al., 2020). These studies are usually done from the point of view that these NLP models might potentially be used to make hiring decisions, which are feared to perpetuate gender biases in future hiring outcomes, especially in the field of software engineering (Blodgett et al., 2020; De-Arteaga et al., 2019; Bolukbasi et al., 2016).

While these biases are important in their own right, other aspects of gender bias are also studied, but typically less often. However, of these other studies, general representational biases are usually tackled, with the ultimate goal of not disproportionately representing different genders in demeaning ways (Nangia et al., 2020; Nadeem et al., 2021; Blodgett et al., 2020). These studies usually focus on a variety gender stereotypes that are chosen in a non-systematic way. Usually, the biases or stereotypes that are chosen are all grouped together under the category "gender bias".

Despite this, we also believe that several gender biases are more prominent in several cultures and that these biases are most expressed in the associated language that is most used in said culture. For example, we will examine gender biases relating to politeness in Japanese and Korean, where politeness is an unavoidable aspect of these languages. Thus, the topic of limited coverage of gender biases has overlap with the current state of English-centric research.

1.2.3 Low-Quality Data

Finally, we note the low data quality in several prominent studies on gender bias. These shortcomings apply mostly to studies that use natural text (as opposed to templates) to evaluate gender bias, such as that of Nangia et al. (2020) and Nadeem et al. (2021). Blodgett et al. (2021) point out many shortcomings of these datasets that rely on annotators to provide examples of stereotypical text. Unfortunately, many problems in these datasets might be simply due to the inattentiveness of the annotators and a non-rigorous proof-reading phase. For example, Blodgett et al. (2021) pointed out an example where the term "women" is compared to the unrelated term "house burglars" for a specific test example.

In addition to these shortcomings, we will also point out that several studies on gender bias in NLP suffer from small datasets, which leads to results that are not as statistically robust. We will attempt to alleviate these statistical shortcomings by proposing statistical measures that lend themselves to being readily and simply interpreted using standard statistical techniques. However, it is best to be specific as to our contributions, which we will summarize in Section 1.4.

1.3 Models and Methods

1.3.1 Strategy for Seeking Gender Biases to Probe

In this thesis we focus on seeking novel methods for probing gender bias. Our general approach was to search for aspects of gender bias that have not been studied, or that have relatively little coverage, and then devise a method to analyze this bias. In particular, in our search for biases, we focused on biases that have a historical basis or that are contemporarily relevant. We also consulted native speakers concerning the relevance of the bias when appropriate.

Additionally, in our search, we also attempted to address the points outlined in Section 1.2, namely we attempted to move beyond English-centric gender bias, extend the coverage of biases considered and strived for retaining high quality data.

With these simple guiding principles, we sought to fill gaps in current research over the course of the time research was conducted for this thesis.

1.3.2 Deep Learning and Transformers

In this thesis we make use of numerous NLP models, which are all deep learning models, specifically models with the transformers architecture (Vaswani et al., 2017). We will give a brief introduction on the core principles behind deep learning and transformers in this subsection.

Deep Learning

Prior to the discussion on deep learning we first have to discuss artificial neural networks for completeness. Artificial neural networks (ANNs) are a system of interconnected units that attempt to learn to approximate a function, f, from data (Goodfellow et al., 2016). The term "neural" comes from ANNs being loosely inspired by neuroscience (Goodfellow et al., 2016). A basic ANN is shown in Figure 1.1. The ANN takes an input $\mathbf{x}$ (where $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ in Figure 1.1) and returns an output $\mathbf{y}$ (where $\mathbf{y} = (y_1, y_2)^T$ in Figure 1.1). A basic ANN, such as that in Figure 1.1, makes use of the hidden units $\mathbf{h} = (h_1, h_2, h_3)$ by linearly combining their inputs via learned weights, $\mathbf{W}$ and $\mathbf{b}$, (e.g. $\mathbf{x}^T\mathbf{W} + \mathbf{b}$ for the hidden weights feeding into the first hidden layer) and subsequently applying a non-linear function g (e.g. the sigma function σ) to the linear combination (i.e. $g(\mathbf{x}^T\mathbf{W} + \mathbf{b})$)) (Goodfellow et al., 2016). The inclusion of a non-linear function is crucial, otherwise the network can only learn linear functions (as the linear combination of linear functions is linear) (Goodfellow et al., 2016; Sutton and Barto, 2018). In this sense, the ANN can essentially be thought of a function f, where $f(\mathbf{x}) = \mathbf{y}$.

The ANN can be trained via a loss function, $\mathcal{L}$, which quantifies how far the predicted output, $\mathbf{y}$, is from the desired output, $\hat{\mathbf{y}}$ (Goodfellow et al., 2016). By taking the derivative of the loss function, with respect to the parameters (i.e. $\nabla\mathcal{L}$), the parameters may be updated (using a method such as stochastic gradient decent, shown in Eq. 1.1 for $\mathbf{W}$ (the same equation can be used for $\mathbf{b}$, by exchanging $\mathbf{W}$ for $\mathbf{b}$), where α is the step size), such that the loss function may be minimized (Goodfellow et al., 2016).

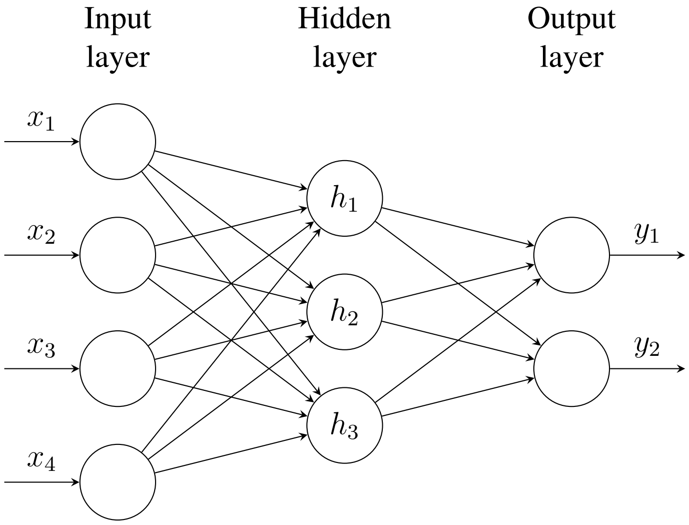


Figure 1.1 – *A basic artificial neural network with four input units and two output units. This network also has one hidden layer.*

$$\mathbf{W} := \mathbf{W} - \alpha \boldsymbol{\nabla}_{\mathbf{W}} \mathcal{L} \tag{1.1}$$

We note that it is possible to learn any continuous function to any desired accuracy on a compact region of the input space with sufficiently many hidden units h_i, for an ANN with a single hidden layer (Cybenko, 1989). However, from practical and theoretical perspectives it is usually easier to model complex functions by increasing the number of hidden *layers*, such that increasingly complicated abstractions of the input may be represented by successive layers of the model (Zeiler and Fergus, 2014; Sutton and Barto, 2018; Bengio, 2009) This is where the term “deep” originates from in deep learning, namely deep learning concerns itself with neural models with numerous hidden layers (Goodfellow et al., 2016; LeCun et al., 2015).

Transformers

Transformers are a type of neural network architecture that achieved state of the art performance and are commonly used in modern NLP models (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2018b). The basic architecture of a transformer is shown in Figure 1.2. The basic architecture consists of an encoder and a decoder. The encoder encodes the symbolic input representation into a continuous representation, which the decoder uses to create an output sequence (Vaswani et al., 2017).

Prior to being fed into the model, the input text is converted into tokens, through the use of a tokenizer, such as Byte Pair Encoding (BPE) (used in GPT-2, which is also built on the transformer architecture) (Sennrich et al., 2016; Radford et al., 2018b) In the embedding layer, the tokens are converted into an embedding, which is a vector containing information regarding the semantic information of the token (Vaswani et al., 2017; Cho et al., 2014). Subsequently, information regarding the position of a token in a sequence is provided by positional encodings, sine and cosine functions sampled at points corresponding to the position of a token in an input sequence (Vaswani et al., 2017). These positional encodings are simply added to the embeddings, which enables encoding additional positional information (Vaswani et al., 2017).

A key innovation behind the transformer architecture is to rely solely on self-attention (Vaswani et al., 2017). Self-attention is a mechanism to relate different positions of a given input sequence to compute a representation of the input (Vaswani et al., 2017). Compared to previous recurrent and convolutional approaches, the use of self-attention drastically reduces the training time as it allows for more parallelized computation and by reducing the path length between long-range dependencies in the network (Vaswani et al., 2017).

This architecture had a great effect on NLP and is the basis of all the models we will discuss in the following sections.

1.3.3 Masked Language Models

Masked Language Models, such as BERT (Devlin et al., 2019), had a large effect on NLP research in the last couple of years. In this discussion, we will focus on BERT, a transformer-based model that was considered state of the art at the time of publication, and still remains a vital baseline in current studies (Rogers et al., 2020)

The BERT architecture is almost identical to the transformers architecture of Vaswani et al. (2017) we discussed earlier (Devlin et al., 2019). A key point of BERT is that it uses bidirectional self-attention, meaning each token attends to context both to its left and to its right (Devlin et al., 2019). This is in contrast to the autoregressive OpenAI GPT transformer models, where the models are constrained, such that each token can only attend to context to its left (Radford et al., 2018a)

There are two training steps in BERT, the first being pre-training, where BERT is trained on unsupervised tasks using a training corpus, and the second being fine-tuning, where the model is trained on one specific task (Devlin et al., 2019). An overview of the BERT training setup is shown in Figure 1.3. During pre-training BERT is trained on two unsupervised tasks. The first is the masked language model objective, where 15 % of input tokens are *masked*, meaning they are re-

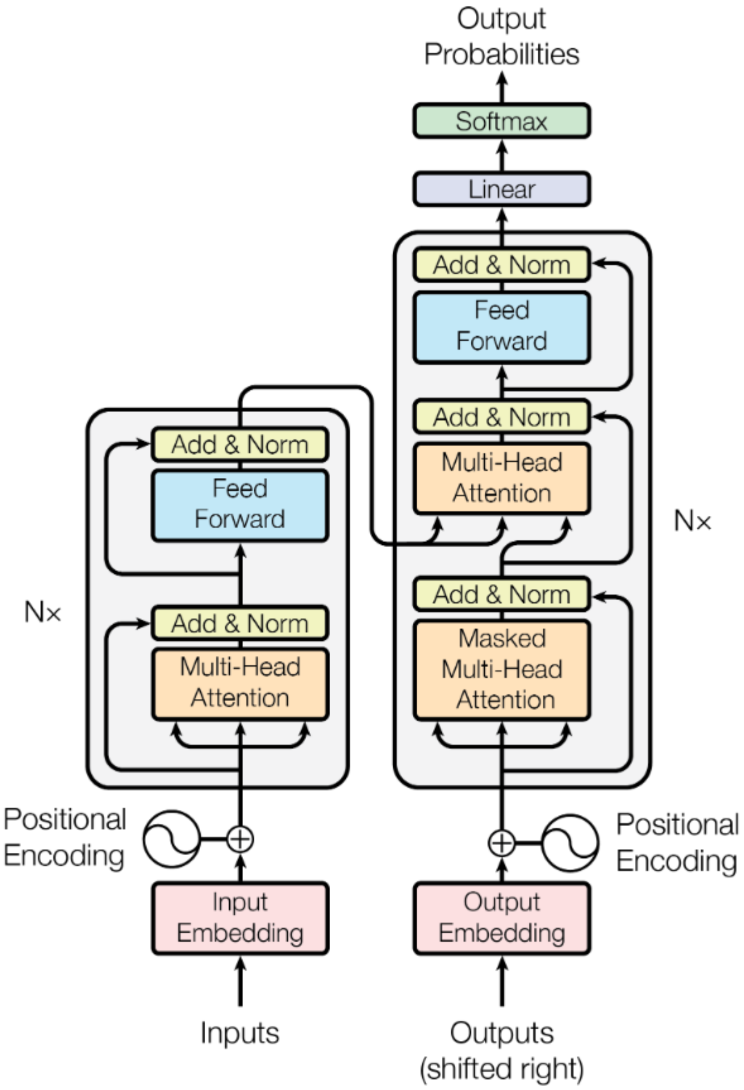


Figure 1.2 – *The basic transformer architecture. The left side of the architecture is the encoder and the right is the decoder. Both the encoder and decoder are stacks of $N = 6$ of the shown units in the original transformer paper (Vaswani et al., 2017). Figure taken from (Vaswani et al., 2017).*

placed with a special token, `[mask]`, and the model has to predict the original token that was there prior to masking (Devlin et al., 2019). This very closely follows the cloze procedure from gestalt psychology, where human participants are asked to complete a sentence template with a missing word (Taylor, 1953; Devlin et al., 2019). The second is the next sentence prediction task, where for a given sentence from the corpus the following sentence (separated by a special `[SEP]` token) is provided 50% of the time and a random sentence from the corpus is selected the other 50 % of the time (Devlin et al., 2019). The model is then trained to identify if the second sentence follows the first in the corpus or if it is a randomly selected unrelated sentence (Devlin et al., 2019).

Fine-tuning is comparatively far less computationally taxing than pre-training as it involves only modifying the inputs and outputs to the task and fine-tuning all the parameters end-to-end (Devlin et al., 2019). For example, as part of the General Language Understanding Evaluation (GLUE) benchmark, the Multi-Genre Natural Language Inference (MNLI) test probes a model's ability to determine if a given sentence (`sentence B`) is an entailment, contradiction or neutral with respect to the first sentence (`sentence A`) (Wang et al., 2018; Williams et al., 2018). To do this, the model's input is provided in such a way that `sentence A` and `sentence B` are separated by the `[SEP]` token and the output layer is modified such that it provides a single classification label (Devlin et al., 2019). Thus, only the input and output have to be changed for fine-tuning.

Besides BERT, other masked language models have also emerged that are quite similar in structure and training but achieved greater performance. For example, RoBERTa achieved greater performance than BERT by improving on several of BERT's hyper parameters (Zhuang et al., 2021). Similarly, ALBERT also improves on BERT by using parameter reduction techniques to reduce memory requirements and by using different loss functions (Lan et al., 2020).

1.3.4 Autoregressive Models

For autoregressive, or generative text models, we will make use of the popular open-source model GPT-2 (Radford et al., 2018b). The autoregressive nature of GPT-2 is enabled by feeding the generated output tokens back in as input for the next tokens, as was discussed in the original transformers paper (Vaswani et al., 2017). GPT-2 is pre-trained on the language modeling objective (Radford et al., 2018a,b). The language modeling objective is to maximize the prediction probability of the next correct token (Radford et al., 2018a,b). Mathematically, the likelihood, L, in Eq. 1.2 is maximized by updating the model parameters Θ (encapsulating all trainable parameters), where u_j is the j^{th} token of a corpus of tokens, and k is the size of the context window of the model (Radford et al., 2018b,a).

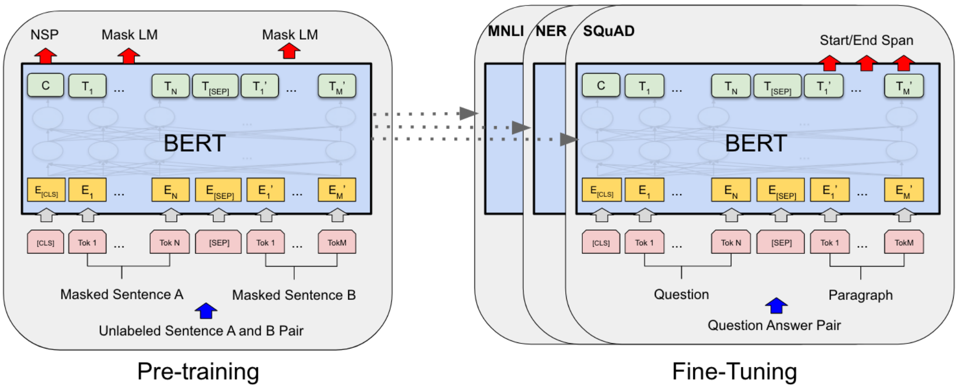


Figure 1.3 – *BERT training setup overview. The model is first pre-trained on unsupervised tasks using a training corpus, and subsequently the pre-trained model is fine-tuned to the desired downstream task (Devlin et al., 2019). Figure taken from (Devlin et al., 2019).*

$$L = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}, \Theta) \tag{1.2}$$

Like the previous models we discussed, the model architecture is based on the transformer architecture of Vaswani et al. (2017). The architecture builds on that of its predecessor, GPT (Radford et al., 2018a), with some slight modifications, including an additional layer normalization (Ba et al., 2016) after the final self-attention block (Radford et al., 2018b).

Finally, for the input representation, the model makes use of Byte-Pair Encoding (BPE) (Radford et al., 2018b; Sennrich et al., 2016). In BPE frequent character sequences are combined into separate tokens to more efficiently make use of the model's capacity (Radford et al., 2018b; Sennrich et al., 2016). BPE was empirically found to combine the performance benefits of word-level-tokenized language models, as well as the generalizability of byte-level models (Radford et al., 2018b).

1.3.5 Sentence Transformers

For classification tasks, we will make use of Sentence Transformers (Reimers and Gurevych, 2019), and classification heads that have been fine-tuned within the SetFit framework, which has been shown to exhibit less variability and to be more sample efficient than other popular fine-tuning techniques (Tunstall et al., 2022). Sentence Transformers, such as Sentence-BERT, make use of a pre-trained transformer model, such as BERT (Devlin et al., 2019) and train it in a Siamese setup.

In the Siamese setup, which was first introduced in Bromley et al. (1993) for detecting forged signatures, a neural network learns to differentiate training samples of dissimilar classes and associate training samples of the same class (Reimers and Gurevych, 2019; Tunstall et al., 2022; Bromley et al., 1993). During training, two identical sub-networks (which are chosen to be the pre-trained BERT model (Devlin et al., 2019) for Sentence-BERT (Reimers and Gurevych, 2019)), are fed different sentences, belonging to the same or different class, and the output representations of the sub-networks are compared with a contrastive loss function, such as the cosine-similarity loss, in the case of SetFit (Tunstall et al., 2022). Note that the two sub-networks have identical weights to ensure that the two sub-networks are identical (Reimers and Gurevych, 2019; Tunstall et al., 2022; Bromley et al., 1993).

One important feature of this contrastive setup for few-shot learning (situations with a few number of labeled training examples, K) is that because during training we require training pairs (as opposed to single training examples) we artificially enlarge the training set (Tunstall et al., 2022). Specifically, the number of unique training pairs that may be chosen from K training samples is simply the binomial coefficient shown in equation Eq. 1.3 (Tunstall et al., 2022).

$$\binom{K}{2} = \frac{K(K-1)}{2} > K, \quad \forall K > 4 \tag{1.3}$$

Using this setup, the sub-networks learn sentence embeddings that may be used to distinguish sentences of different classes (Tunstall et al., 2022). The final step in SetFit is to train a classification head that learns the associated class for each of the different embeddings, for this a simple logistic regression model is used (Tunstall et al., 2022).

1.3.6 Image Captioning Models

Image captioning models, as their name suggests, are designed to predict a possible caption that could be used to describe an image (Wang et al., 2022). In our case, we made use of Microsoft's popular open-source model, GIT (Wang et al., 2022).

The model consists of two main components, an image encoder and a text decoder (Wang et al., 2022). The image encoder is a ViT (Vision Transformer) model and only takes an image as input (Dosovitskiy et al., 2021; Wang et al., 2022). This encoder was pre-trained using CLIP (Contrastive Language-Image Pre-Training), where the image encoder was jointly trained with a text encoder to distinguish between different images and their associated captions (Radford et al., 2021; Wang et al., 2022). Unlike in the previous text models where tokens are inputs to the models, in image models, such as ViT the RGB pixel values of the image are the inputs (Dosovitskiy et al., 2021).

The text decoder is trained similarly to GPT-2, with the only exception being that the decoder is also conditioned on the encoded representation of the image, I (Wang et al., 2022). In other words, the text encoder is trained to maximize the likelihood in Eq. 1.4, where $\{u_j\}$ are the tokens of the text caption associated with the image (Wang et al., 2022).

$$L = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}, I, \Theta) \tag{1.4}$$

1.4 Contributions

1.4.1 Project 1 - Multilingual Dataset and Information-Theoretic Measure

In the first project we moved beyond the English-centric nature of current bias research and formalize a method to measure bias that is applicable to all languages.

We will first discuss the dataset. Building off of the research of Nangia et al. (2020), we built a multilingual gender bias dataset using sentence pairs. The sentence pairs are gender-swapped versions of each other, where one sentence is considered to be more stereotypical than the other, as determined by the annotators in the original CrowS-Pairs study (Nangia et al., 2020). Consider example (1), in this case the sentence “Robert is a pilot.” would be considered more stereotypical than “Olivia is a pilot.” as dictated by social gender stereotypes.

(1) a. **Robert** is a pilot.
b. **Olivia** is a pilot.

A key point of these pairs of sentences is that they may be used to compare gender differences across any language, so long as we can find gendered names that can replace the names “Robert” and “Olivia” with names that are common in the target language. The parts of the sentences that are compared are ones that remain the same after gender swapping, in the example here, that would be “is a pilot.”. The reason for using names stems from the fact that many languages, such as Turkish or Swahili, make use of gender-neutral pronouns or they don’t have any gendered words, so we cannot rely on pronouns such as the English pronouns “he” or “she” to identify gender. For this project we created a cleaned version of the CrowS-Pairs (Nangia et al., 2020) dataset, which we then translated to nine additional languages with the help of native speakers.

A second key contribution of this project was proposing the S_{JSD} bias measure. Inspired by information theory, this measure aims to provide a score of how biased a given sentence pair is, represented by a real number that can take on a range of values. This contrasts to previous approaches such as that of Nangia et al. (2020) and Nadeem et al. (2021), which simply assigned a binary score of 1 if a model considered the more stereotypical sentence more likely than the non-stereotypical, and 0 otherwise. The measure is shown in Eq. 1.5, where $JSD(P||G)$ is the Jensen-Shannon distance between the probability distributions P and G (Lin, 1991). In our case, P_{more} and P_{less} are the model's output probability distribution of the more and less stereotypical sentences, respectively. Additionally, the probability distribution G is a one-hot distribution identifying the correct tokens of the parts of the sentences that overlap. We demonstrate this measure is more suitable to use in situations where we have smaller datasets, as is the case in numerous bias datasets, where it is expensive to obtain high-quality structured data.

$$S_{\text{JSD}} = \sqrt{\text{JSD}(P_{\text{more}}||G)} - \sqrt{\text{JSD}(P_{\text{less}}||G)} \tag{1.5}$$

1.4.2 Project 2 - Politeness Stereotypes and Attack Vectors

In the second project we moved beyond the common notions of gender bias, such as occupation bias, that are pervasive in the NLP community, to focus on a more subtle form of bias, namely politeness bias and demonstrate how these biases can be harmful in downstream tasks.

In the first part of the project, we demonstrated that female speech is characterized by more non-formal polite speech, while male speech is comparatively more associated with formal and rough speech. To do this, we made use of Korean and Japanese politeness levels, which are encapsulated in verbs. For example, consider example (2), which compares two different politeness levels of the Japanese "to study". The first instance is more appropriate among family members while the second is more suitable among strangers and when one wishes to speak more politely about a topic (Eri et al., 2011). Included in our study of politeness levels, we also compare the effect of honorifics. To probe for biases, we make use of specially designed templates with the aid of native speakers which we feed into popular MLMs with Korean or Japanese support.

(2) a. 勉強する。 (benkyou suru)
b. 勉強します。 (benkyou shimasu)

In the second part of the project, we demonstrated that learned associations between gender and politeness levels can influence downstream performance. Specifically, we investigated cyber bullying detection systems for gender differences in cyber bullying detection, when the system is provided an example of hate speech supplemented with information as to who the hate speech is directed at, using different politeness levels. This makes our experiments similar to Gröndahl et al. (2018), who demonstrated that appending the word “love” to the input of hate speech detection systems renders them ineffective. In our experiments we demonstrated that the use of honorifics makes cyber bullying systems effective when male speakers are being referred to but ineffective when referring to female speakers. Based on our results we believe looking into biases beyond the typical occupation biases or general gender stereotypes in languages beyond English is a fruitful field of future research.

1.4.3 Project 3 - Textual and Visual Encapsulation of Bias in Emoji

In the third project we investigated harmful language generation, triggered by textual and visual representations of emoji.

To investigate biases regarding textual representations of emoji, we limited ourselves to Unicode representations of emoji. For example, the emoji for a smile (🙂) will simply be represented as its Unicode code point, namely U+1F642 (Unicode Consortium, 2022). We then made use of HONEST, a method for probing generative models for harmful output, as outlined in the research of Nozza et al. (2021), where we used the same templates but prepended with emoji Unicode code points. Using GPT-2 (Radford et al., 2018b), we demonstrated similar rates and patterns of harmful word detection as found by Nozza et al. (2021), where female person references evoked higher rates of harmful word detections. However, based on our experiments the use of emoji Unicode code points does not seem to greatly influence the gender imbalances of harmful sentence completions, beyond the imbalances already covered by Nozza et al. (2021).

In similar spirit, we also experimented with supplying visual information to image captioning models. For this, we probed Microsoft’s GIT model, fine-tuned on COCO (Wang et al., 2022; Lin et al., 2015). We provided images of common emoji, such as the wink emoji (😉), as exemplified in Figure 1.4, together with the beginning of a sentence including a gender identity term, which is to be completed by the model. For the beginning of sentences, we once again make use of HONEST (Nozza et al., 2021), but this time without any modification regarding textual input. Thus, the visual and textual inputs are separate and only interact with each other through the learned parameters of the model.

Evaluating the results, we find that heart-shaped emoji (♥(heart suit), 🖤(black heart) and ❤️(red heart)) lead to more harmful text generation for female identity terms, while less emotive emoji, such as the smile and frown emoji (🙂and 🙁)

Figure 1.4 – *An example image for the visual representation experiments. Here the Apple wink emoji is centered on a white background.*

invoke more harmful text generation for male identity terms. The reason for these gender differences are not known and are assumed to be a result of learned gender stereotypes in the training set, but more tests would be needed to identify the reason.

1.5 Summary

To summarize, in this thesis we will outline three projects that aim to broaden the breadth of current gender bias research in NLP. The first project aimed to build a dataset that investigates languages beyond English and that may be simply extended to any language of choice. In addition, we proposed a new analytical bias measure that may be used to evaluate these biases, given the model's prediction probabilities. In the second project we demonstrated that learned gender stereotypes regarding politeness may bleed into cyber bullying detection systems, which may disproportionately fail to protect women if the system is attacked with honorifics. Finally, the third project showcased that visual representations of emoji may evoke harmful text generation that disproportionately affects different genders, depending on the emoji choice.

Chapter 2

An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models

Everything not saved will be lost.

Nintendo quit screen

In this project we tackle gender biases in languages beyond English. While NLP bias research traditionally centered on English, we propose a technique and dataset that may be simply extended to any language by making use of gender-swapped sentence pairs.

Specifically, our contributions are as follows:

- Created a dataset for probing gender biases across languages
- Developed a novel information-theoretic measure that aims to distinguish strong and weak biases in NLP models.

This work was published in Findings of NAACL 2022 (Steinborn et al., 2022). The authors were Victor Steinborn, Philipp Dufter, Haris Jabbar, Hinrich Schütze.

2.1 Introduction to Multilingual Bias

Pre-trained language models (PLMs) have greatly benefited NLP (Raffel et al., 2020; Peters et al., 2018; Devlin et al., 2019; Zhuang et al., 2021). However, commonly used PLMs such as BERT have been shown to encapsulate social biases, including those relating to gender and race (Kurita et al., 2019; Nadeem et al., 2021; Nangia et al., 2020). The general consensus is that these biases are learned from the statistical distributional co-occurrence of words relating to a group (such as terms relating to men or women) with a context in which that group is often mentioned in corpora (Bolukbasi et al., 2016; Webster et al., 2021). For example, “doctor” may co-occur with “man” more often than with “woman”, leading to an internal representation in the model where a gender-neutral concept, such as being a doctor, is more closely associated with male-related terms than with female-related terms (Bolukbasi et al., 2016).

In this work we tackle these types of binary stereotypical representational gender bias (henceforth simply “gender bias”) in MLMs in a multilingual setting. We argue that unbiased models should not give different prediction probabilities for tokens that remain unchanged after changing the gender of a person the text refers to. Based on this, we propose a multilingual approach to study gender bias in MLMs, outlined in Figure 2.1, which, to the best of our knowledge, can in principle be extended to any natural language.

The importance of developing AI systems that are mindful of different societal groups, such as people of different genders, is a topic much discussed in the area of

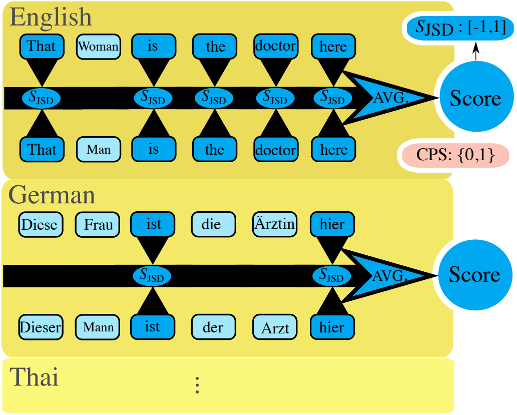


Figure 2.1 – *Following (Nangia et al., 2020), we assess multilingual gender bias in MLMs by matching gender-specific tokens (light blue) in the context of non-gender-specific tokens (dark blue) in sentence pairs. We develop a methodology for creating sentence pairs that we argue is applicable across languages, in contrast to prior work. We mask unchanged tokens one at a time and calculate S_{JSD}, a novel information-theoretic bias measure whose sentence-level average we show to be better behaved than competing measures.*

fairness research in NLP (Blodgett et al., 2020). However, a shortfall of this area is its almost exclusive focus on English. As far as we are aware, ours is the first study to attempt to create a truly multilingual approach to study gender bias in language models. Previous multilingual approaches were largely limited to sentences with fixed templates and grammar structures, which heavily constrains the range of languages that may be studied with a given template (González et al., 2020). Our approach builds on (Nangia et al., 2020) and attempts to study natural sentences by comparing a pair of sentences that differ only by the gender of persons mentioned, a process which we will refer to as *gender swapping*.

To illustrate the problem of using templates, consider the following sentence pair and its German translation.

(1) a. **He** is the doctor here.
b. **She** is the doctor here.

(2) a. **Er** ist **der Arzt** hier.
b. **Sie** ist **die Ärztin** hier.

In German the only parts that remain the same are “ist” and “hier” under gender swapping, as the German word for the profession “doctor” and its associated definite article change form depending on the gender of the person. Thus, template structures developed for English of the form

(3) [*person*] is the [*profession*] here

have to be heavily modified and constrained to create grammatically correct sentences in German. The problem is exacerbated with multilingual studies, where appropriate templates need to be decided upon for each language.

We take inspiration from CrowS-Pairs (CPS) (Nangia et al., 2020), which studies pairs of crowd-sourced sentences, for a range of social biases. It includes gender-swapped pairs for the diagnosis of gender bias. However, we found that we cannot simply translate CPS into other languages. The main problem is that English pronouns are clear indicators of gender – at least of binary gender, which we focus on in this project. But this clear indication gets lost in translation for languages that have gender-neutral pronouns like Finnish and those that predominantly use null pronouns like Thai.[1] We could mandate that only words with "gender-inherent" meaning like "mother", "wife" and "sister" are used, but that would exclude many topics that we need to cover in a good diagnostic dataset, e.g., work life and sports.

The solution we propose is to simply use names to indicate gender. Our assumption here is that all languages have words for names and that there are two subsets of names that can only have female and male referents. Note that there are certainly "unisex" names, i.e., names that can refer to both men and women, even in English ("Jess", "Leslie"). But as far as we know no language has been discovered in which there are no "monosex" names, i.e., names that can refer to only one gender. We rely on such monosex names to construct sentence pairs.

In English, we select a few frequent male and female names and only use them for the English dataset. Before translating the sentence pairs into another language, we first identify corresponding frequent male and female names in the target language. The translators are then instructed to only use those names. This methodology should be applicable universally, so that we can construct a multilingual gender bias resource for any set of languages. Additionally, the CPS dataset has been edited to heed the recommendations of (Blodgett et al., 2021). We initially translated the modified CPS dataset into German, Indonesian, Thai and Finnish. Subsequently, after publication, we extended the translations to Arabic, French, Korean, Vietnamese and Chinese. A more detailed description will be given in Section 2.3.1.[2]

The second contribution of this project is S_{JSD}, a novel measure based on the

[1] The English sentence "she ate it" is simply expressed as "ate" in many "pro-drop" languages as long as subject and object of "ate" are clear from context.

[2] Blodgett et al. (2021) argue against using names for race. Their arguments do not apply to gender. See Section 2.3.1.

Jensen–Shannon divergence (Lin, 1991), to test MLMs for social biases by using sentence pairs that capture a binary contrast between two groups. The measure used in CPS (see Section 2.3.2) makes use of a binary decision process, which has the effect of removing information of the probability values from the MLM, which we show reduces the measure's predictive power. Our motivation for introducing S_{JSD} is to retain as much information from the MLM output probabilities as possible in our final reported score in order to make effective use of the limited amount of human-translated sentences that are available.

Thus, our contributions are (1) developing a method for creating multilingual datasets for diagnosing gender bias in language models that is applicable across the diverse set of human languages, (2) applying this method, taking the CPS dataset (Nangia et al., 2020) as a starting point, and creating a multilingual gender bias diagnosis dataset for ten different languages, (3) proposing the S_{JSD} measure, which retains information regarding the numeric output probabilities of MLMs.

2.2 Related Work

Given this work focuses on multilingual methods to measure gender bias in MLMs, this discussion will focus on evaluation measures and techniques; a thorough discussion of debiasing methods is beyond the scope of this project.

2.2.1 Bias Measures in MLMs

Recently, pre-trained masked language models, such as BERT (Devlin et al., 2019), have significantly gained in popularity, which in turn has led to numerous studies analyzing their behavior, including their encapsulation and reproduction of social bias. Prior to the emergence of these models however, it was already well known that NLP models can learn social biases from corpora, as exemplified in work by Bolukbasi et al. (2016) who demonstrated that word embeddings encapsulate societal gender biases. Subsequently, further tests, such as the word embedding association test (WEAT) by (Caliskan et al., 2017), demonstrated that word embeddings also encapsulate other biases, including racial biases. May et al. (2019) extended WEAT to sentence encoders, including BERT with the sentence encoder association test (SEAT), to study sentence-level social biases in these models using template constructed sentences. However, the results of this study were not conclusive, and Kurita et al. (2019) showed that the cosine-based methods used in WEAT and SEAT are not appropriate for contextualized embeddings, and instead use a scoring method based on the prediction probability of an attribute given a target in template sentences.

The evaluation method used in the StereoSet (Nadeem et al., 2021) was inspired by SEAT while CPS (Nangia et al., 2020) uses pseudo-log-likelihood MLM scoring (Salazar et al., 2020). A contribution of CPS and StereoSet is to provide techniques that evaluate natural sentences instead of simple templates. Nonetheless, Kaneko and Bollegala (2021) criticize CPS and StereoSet for their evaluation measures, arguing that the act of masking tokens results in a systematic overestimate in measured biases. However, they also describe this effect as systematic, and thus we would expect systematic trends in bias scores between models to remain conserved when masking tokens.

2.2.2 Multilingual Studies of Bias in MLMs

As far as we are aware, there are no studies that have attempted to develop a multilingual method to test for gender bias in MLMs without template structures. However, there are several that studied a well-defined set of languages. For example, González et al. (2020) constructed sentence templates for languages with type B reflexivization (including Swedish and Russian), which can be used to construct challenge datasets to measure gender bias. Bartl et al. (2020) also constructed templates to study biases in German and English BERT models, but sometimes a different form of a template has to be used depending on the gender of a mentioned person. Liang et al. (2020) examined the case of English and Chinese using templates while focusing on the cross-lingual transfer of removing biases in Chinese using English training data.

2.2.3 Bias From a Social Science Perspective

A critical survey of 146 NLP papers by (Blodgett et al., 2020) outlines common pitfalls in NLP research, including the CPS study, when attempting to study social bias. We attempt to take into account in this work.

2.3 Methodology

2.3.1 Dataset

A major obstacle in transferring existing techniques to measure gender bias in languages beyond English is in adapting methods to the target language's gender agreement system. Methods intended to measure gender bias in MLMs often rely on fixed sentence templates, where predefined words are inserted that are intended to test some aspect of bias, such as occupational gender bias (e.g., (Kurita et al., 2019; Webster et al., 2021)). While these template structures can be modified and applied to a range of languages, once a template is chosen, the range of languages that can be studied is restricted (González et al., 2020).

Thus, to design a multilingual approach to gender bias, we want to move beyond the rigid artificial sentence structures that result from using templates. We also speculate that moving away from rigid sentence structures allows us to probe the language model more deeply for biases. It may be possible that superficially debiased language models can perform well on certain bias evaluation tasks that use templates, similar to the situation for linearly debiased word embeddings that perform well on some bias measures but still encapsulate significant distributional biases (Gonen and Goldberg, 2019).

Two evaluation datasets that go beyond templates are StereoSet (Nadeem et al., 2021) and CPS (Nangia et al., 2020). One important difference between their methods is the masking pattern. While StereoSet's context association test masks words that may be gendered in a different language (e.g., adjectives in Spanish), CPS consists of pairs of sentences and only masks tokens that are shared by the two sentences. Here we will only consider the CPS dataset, which also marks which of the two sentences is more stereotypical (Nangia et al., 2020).

For our dataset, we consider sentence pairs where people of the male and female gender are being contrasted, for example:

(4) a. **He** is a pilot
b. **She** is a pilot.

For this example, we assume each word is a separate token. The unmodified tokens common to both sentences are: "is", "a", "pilot". For each sentence, the unmodified tokens form a set U and the remaining modified tokens a set M ("**He**" for example (4), sentence a, for example). Thus, for each sentence, the set of all tokens is the union of U and M.

Connecting to multilinguality, as far as we are aware, for sufficiently long and complex sentences, when swapping the gender of a person reference in a sentence there remain sections of the sentence that remain unchanged in any natural language. From this observation, we found the masking pattern CPS implements to be appropriate for multiple languages and thus the sentences labeled with the "gender" tag in the CPS dataset were selected as the basis for subsequent translations.

It is worth mentioning that the CPS dataset recently received criticism for lacking clear explanations of what types of social biases are being measured (Blodgett et al., 2021). For this reason the selected CPS sentences have been minimally modified to be mindful of the pitfalls outlined in (Blodgett et al., 2021). Some sentences were omitted because the contrasted groups were unrelated to the stated "gender" label, such as for sentences that contrasted two racial groups instead.

We will now outline the modifications of the CPS dataset for this study.

First, we ensured each sentence only compares binary gender. Non-binary gender adds a level of complexity in the multilingual context that we decided to leave for future work. We also removed sentences that compare clothing items, most likely intended as a proxy for gender. Clothing items and their significance differ across cultures, so such sentences are difficult to translate.

Second, for sentences that only used a pronoun to identify gender, we exchanged the pronoun with a common name that is stereotypically associated with one gender in the English dataset. Subsequently, when translating the English dataset into other languages, the names were exchanged for others that are common gendered names in the target language. We limited the number of names in the English dataset to four to simplify the subsequent translation process. Names were introduced because many languages do not have gendered pronouns, and thus information relating to gender may be lost in translation. For example, a typical translation of (4) into Indonesian results in two identical sentences, which makes the sentence pair useless for Indonesian. Using names as a proxy for identifying a social group is discouraged in (Blodgett et al., 2021) for race bias, but using stereotypically gendered names as a proxy for binary gender seems unproblematic to us. For example, whereas names only indirectly and ambiguously identify race (at least in English), we can easily find names that are "monosex", i.e., names that can only have either male or female referents. Thus, we would modify example (4) as follows for our dataset:

(5) a. **Robert** is a pilot
b. **Olivia** is a pilot.

Finally, we removed sentences that did not correctly isolate a stereotype, an issue noted in the original paper (Nangia et al., 2020).

In this work we investigate binary gender stereotypes as a representational harm across languages, to use the terminology of (Blodgett et al., 2020). The CPS dataset was created by US crowdworkers (Nangia et al., 2020). We make the assumption that most aspects of gender bias should be part of a diagnostic test across languages and cultures. For example, the associations of "doctor" with "male" or of "childcare" with "female" are biases that most cultures are at risk for. So we should test whether our language models exhibit these biases for all cultures. There probably are aspects of gender bias that are relevant to only a few cultures (e.g., maybe the association of "being eligible to drive a car" with "male"). We stress the importance of investigating gender bias multilingually. Given that our study is the first to do this, we feel justified to leave the issue of how to comprehensively test for all aspects of bias in gender diagnosis to future work.

	Ar	De	En	Fi	Fr	Id	Ko	Th	Vi	Zh
#w	4949	5470	5548	4151	6136	4790	10019	6693	7623	7430
#w/s	12	13	13	10	14	11	24	16	18	18

Table 2.1 – *Our multilingual bias diagnosis dataset consists of 212 sentence pairs in five languages. The table gives total number of words (#w) and words per sentences (#w/s) for each language. Thai was tokenized with Deepcut (Kittinaradorn et al., 2019), Korean with Kiwi (Lee, 2022) and Chinese with Jieba (Sun, 2020).*

Note that we do not make the assumption that gender bias is the same across languages! If "childcare" is strongly associated with "female" in (the training corpus of) language A, but not in (the training corpus of) language B, then (assuming we use models that pick up bias from their training corpora) our methodology will find less gender bias for language B – and this would be the intended result of our work.

For the translations, we hired translators to translate the modified English dataset into their native language. Translators were paid an agreed upon amount above the minimum wage in their respective country of residence and were informed of the intended use of their translations. Each translator was provided an instruction sheet, which exemplifies the translation process of CPS sentence pairs from English to German. The translation instructions can be found in Section 2.8 as supplementary material and the target languages of the translations were German (De), Finnish (Fi), Indonesian (Id) and Thai (Th) initially. The translations were later extended to include Arabic (Ar), French (Fr), Korean (Ko), Vietnamese (Vi) and Chinese (Zh) after the publication of the associated paper. We chose these languages to cover different language families and because translators for them were easily available to us.

An overview of the metadata of the edited and initial translated dataset is given in Table 2.1.

2.3.2 Bias Measure

Our aim is to create a bias measure that can retain meaningful information from the model output that is relevant for detecting multilingual gender bias. Before introducing our proposed measure, we will go over the CPS measure (Nangia et al., 2020).

CrowS-Pairs Measure

Given is a pair of gender-swapped sentences. One sentence is considered socially more stereotypical than the other by the annotators in the CPS study (Nangia et al., 2020). The set of tokens that are shared (resp. are not shared, i.e., modified) between the two sentences is denoted as U (resp. M) – see Section 2.3.1. For each sentence the tokens in U are masked one at a time. Each time a token is masked, the sentence is passed through the model and the model output probabilities are obtained. Following Nangia et al. (2020)'s notation, we denote the output probability of the model for the i^{th} correct token under the mask $u_{G,i} \in U$ in the more stereotypical sentence as $P_{\text{more}}(u_{G,i}) \equiv P(u_i|U_{\setminus u_i}, M, \theta)$, where M are the unique tokens in the more stereotypical sentence and θ are the model parameters. The output probability for the other sentence P_{less} is defined analogously.

The score for a sentence in the pair is its pseudo-log-likelihood, calculated as the sum of $\log P(u_{G,i})$ over all u in U where P is either P_{more} or P_{less}. The sentence pair is assigned a binary score of 1 (resp. 0) if the more stereotypical sentence has a larger (resp. smaller) score. A possible advantage of this binarization is that the numerical value of the pseudo-log-likelihood cannot be interpreted (hence "pseudo") (Nangia et al., 2020; Salazar et al., 2020), so one can only rely on the comparison of the scores, not on their absolute values. The final score is the percentage of sentences that have been assigned a score of 1.

According to (Nangia et al., 2020), an ideal unbiased model would achieve a score of 50 on a dataset. However, it is important to keep in mind that each sentence pair contributes with equal weight to the final score, due to binarization. Consider as an example a language in which a small part of the sentence pairs are diagnosed as extremely biased, but most sentence pairs do not show bias, so their final score will be randomly 0 or 1. In such a case, CPS does not distinguish strong bias from weak bias and sentence pairs that are not biased contribute noise to the final measure. Hence, unusually biased behavior of the model may not be effectively captured by the measure, and in order to obtain meaningful results a large number of human-annotated sentence pairs is required.

To make the connection to dataset size, if we ignore the internal mechanisms of the model and for simplicity assume that a biased model has a fixed probability, of say $p = 0.55$, to assign a binary score of 1, then this may be modeled as a Bernoulli process (Papoulis and Pillai, 2002). For such a model and for a set of $n = 200$ sentence pairs, the expected dataset score is 55 and the standard error 3.5 (since the standard error is $\sim \frac{1}{\sqrt{n}}$ for Bernoulli). Thus, the CPS measure must rely on a large number of sentence pairs to obtain statistically meaningful results because of the binary decision process that disregards information regarding the extent of the discrepancy between S_{more} and S_{less}. The measures of Nadeem et al. (2021) in StereoSet and of Kaneko and Bollegala (2021) also employ binarization and therefore do not make efficient use of the available data to measure bias.

The Proposed S_{JSD} measure

Our goal in developing the S_{JSD} measure was to create a theoretically well founded measure that retains information regarding MLM output probabilities, avoiding the binary decision process in CPS. This is especially important for our study, where we had limited resources to create the translated dataset.

The S_{JSD} measure is based on the Jensen-Shannon divergence (Lin, 1991), a quantity bounded to the range $[0, 1]$, that measures the similarity between two probability distributions, P and Q, defined as follows:

$$\text{JSD}(P||Q) = H\left(\frac{P+Q}{2}\right) - \frac{H(P) + H(Q)}{2} \tag{2.1}$$

where H is the Shannon entropy (Shannon, 1948).

If P and Q are unrelated and share no overlap $\text{JSD}(P||Q) = 1$ and if they are the same distribution (maximum overlap) $\text{JSD}(P||Q) = 0$. The square root of the Jensen-Shannon divergence (also known as the Jensen–Shannon distance) is a metric, i.e., it satisfies a range of properties intuitive to measures of distance, including the triangle inequality (Endres and Schindelin, 2003). Define the *gold distribution* as a one-hot distribution G that identifies the correct token under the mask. We then define our measure S_{JSD} as the difference of two distances: the Jensen–Shannon distance between P_{more} (resp. P_{less}) and the gold distribution:

$$S_{\text{JSD}} = \sqrt{\text{JSD}(P_{\text{more}}||G)} - \sqrt{\text{JSD}(P_{\text{less}}||G)} \tag{2.2}$$

This definition may also be expressed purely in terms of the model output probability for the token under the mask $P_{\text{more/less}}(u_G)$, as $\text{JSD}(P||G)$ may be expressed in the form shown in Eq. 2.3 for any distribution P. Thus only human annotated text is evaluated.

$$\text{JSD}(P||G) = \frac{1}{2}(P_G \log_2(P_G) - (P_G + 1)\log_2(P_G + 1) + 2), \quad P(u_G) \equiv P_G \tag{2.3}$$

The quantity S_{JSD} is also bound to the range $[-1, 1]$, which limits the effect of outliers. The theoretically ideal non-biased model should yield a value of 0 for S_{JSD} when the distance of P_{more} to G is equal to the distance of P_{less} to G. When P_{more} is closer to G than P_{less}, we take this as a sign of bias for the stereotypical sentence, thus we expect biased models to systematically generate negative S_{JSD} scores.

To generate a score for a sentence pair, we take the average of S_{JSD} scores. Subsequently, for the score on the dataset we simply take the average of the sentence scores.

Model	Multilingual	Parameters
mBERT	yes	178M
xlmR (base)	yes	278M
xlmR (large)	yes	560M
BERT (uncased)	no	110M
RoBERTa	no	355M
ALBERT	no	206M

Table 2.2 – *Details of models used in this study.*

Error Analysis

For an analysis of the error of the reported score on the dataset, we bootstrap the sentence scores to determine an estimate for the standard error using SciPy (Efron and Tibshirani, 1993; Virtanen et al., 2020). For CPS we achieve this by bootstrapping the binary sentence scores.

2.4 Experiments

For our experiments we make use of the Transformers library (Wolf et al., 2020). We use two multilingual models, multilingual BERT (mBERT) (Devlin et al., 2019), trained on Wikipedia, and base xlm-RoBERTa (xlmR) (Conneau et al., 2020), trained on Wikipedia and filtered CommonCrawl data from the internet (Wenzek et al., 2020). We choose xlmR as it has been shown to significantly outperform mBERT on numerous cross-lingual tasks (Conneau et al., 2020). As of this writing, xlmR seems to be the best performing multilingual model in the Transformers library (Wolf et al., 2020; Conneau et al., 2020). The two models differ in training data by the CommonCrawl, which we assume to be more of a source of bias than Wikipedia, based on the results of the CPS study. Nangia et al. (2020) found RoBERTa, trained on Wikipedia and the CommonCrawl, among other datasets (Zhuang et al., 2021), to generally have higher bias scores, compared to BERT (Devlin et al., 2019), although this was not true for gender bias (Nangia et al., 2020).

We run the two models on our translated datasets and calculate CPS and S_{JSD} scores. Running a model on a single language using an Intel Xeon Processor E5-2680 v2 takes roughly 15 minutes.

We also test S_{JSD} on the models and dataset used in the CPS study (Nangia et al., 2020).

Finally, we test the effect of model size on the scores by comparing the large and base xlm-RoBERTa models. See Table 2.2 for a list of all models used.

Model	Lang.	$S_{\text{JSD}} \times 10^{-3}$	CPS	B.S_{JSD}
mBERT	En	$-0.05_{\pm 1}$	$57_{\pm 3}$	$50_{\pm 3}$
xlmR	En	$-1_{\pm 2}$	$62_{\pm 3}$	$54_{\pm 3}$
mBERT	De	$-1_{\pm 2}$	$57_{\pm 3}$	$55_{\pm 3}$
xlmR	De	$-2_{\pm 2}$	$51_{\pm 3}$	$50_{\pm 3}$
mBERT	Id	$-3_{\pm 1}$	$46_{\pm 3}$	$51_{\pm 3}$
xlmR	Id	$-4_{\pm 2}$	$51_{\pm 3}$	$54_{\pm 3}$
mBERT	Th	$-4_{\pm 2}$	$60_{\pm 3}$	$60_{\pm 3}$
xlmR	Th	$-4_{\pm 2}$	$57_{\pm 3}$	$57_{\pm 3}$
mBERT	Fi	$-0.2_{\pm 2}$	$44_{\pm 3}$	$50_{\pm 3}$
xlmR	Fi	$-3_{\pm 2}$	$51_{\pm 3}$	$53_{\pm 3}$
mBERT	Fr	$-3_{\pm 1}$	$58_{\pm 3}$	$58_{\pm 3}$
xlmR	Fr	$-5_{\pm 2}$	$57_{\pm 3}$	$56_{\pm 3}$
mBERT	Zh	$+0.6_{\pm 1}$	$54_{\pm 3}$	$50_{\pm 3}$
xlmR	Zh	$-0.4_{\pm 1}$	$53_{\pm 3}$	$48_{\pm 3}$
mBERT	Vi	$-1_{\pm 1}$	$51_{\pm 3}$	$52_{\pm 3}$
xlmR	Vi	$+0.3_{\pm 1}$	$51_{\pm 3}$	$50_{\pm 3}$
mBERT	Ko	$+0.7_{\pm 0.8}$	$42_{\pm 3}$	$49_{\pm 3}$
xlmR	Ko	$+0.7_{\pm 0.9}$	$56_{\pm 3}$	$49_{\pm 3}$
mBERT	Ar	$-1_{\pm 3}$	$50_{\pm 3}$	$55_{\pm 3}$
xlmR	Ar	$+1_{\pm 4}$	$48_{\pm 3}$	$50_{\pm 3}$

Table 2.3 – *CPS and S_{JSD} scores and standard errors on our multilingual bias diagnosis dataset for all languages. The S_{JSD} scores systematically identify the stereotypical sentence as indicated by the negative scores (with the exception of Arabic for xlmR). Some CPS scores are below 50, indicating the measure cannot capture the stereotypical behavior of the model for this dataset. The binarized version of S_{JSD} (B.S_{JSD}) is also shown to illustrate the effect of binarization. B.S_{JSD} reports scores of 50 or below in four cases where S_{JSD} is negative, suggesting that binarization reduces the predictive power of the measure.*

2.5 Results and Analysis

Table 2.3 shows results for CPS and S_{JSD} on the entire multilingual dataset. We observe that the CPS measure reports scores well under 50 for multiple languages. This goes against the intuition that MLMs learn stereotypical associations from data: it wrongly suggests that male stereotypes are associated with women and female stereotypes with men. We suspect this behavior of CPS comes from the binary decision problem outlined in Section 2.3.2, which is especially relevant for smaller datasets.

A first indication to suspect that we might be in this regime is that the CPS standard errors are close in value to the estimated standard errors assuming a Bernoulli process, as discussed in Section 2.3.2. We can also observe a cluster-

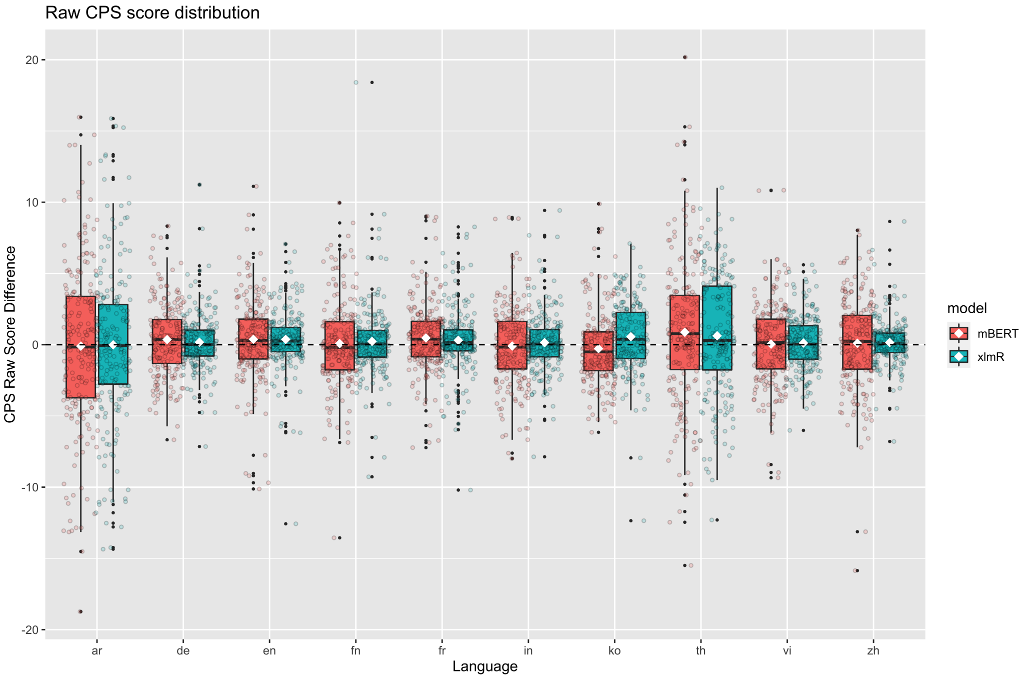


Figure 2.2 – *The difference $S_{more} - S_{less}$ for our entire multilingual bias diagnosis dataset. The white points mark the averages and the box and whiskers plots mark the quartiles. Most of the scores cluster around the decision boundary denoted by the horizontal dotted line.*

ing of CPS sentence scores, before binarization, around the decision boundary in Figure 2.2, indicating that slight variations in bias scores can substantially change the CPS score. Furthermore, the effect of binarizing S_{JSD} (i.e., following the CPS method but replacing $\log P_{\text{more}}(u_{G,i})$ with the JSD distance to the gold token) is shown in Table 2.3. These binarized S_{JSD} scores fail to detect bias by yielding scores of 50 or below in four cases – whereas the S_{JSD} score predicts bias as expected. The drawback of using CPS is especially apparent for Korean, where the S_{JSD} hovers around zero, whereas for CPS the scores fluctuate strongly around zero. All this, coupled with the discussion in Section 2.3.2, reinforces our argument that binarization harms measure performance and that S_{JSD} is numerically more suitable and theoretically justified as a measure compared to CPS, especially on smaller datasets. Note that we did not unbinarize CPS scores as they have no clear statistical interpretation (Nangia et al., 2020; Salazar et al., 2020); see discussion in Section 2.3.2.

Table 2.3 shows that S_{JSD} has negative values, i.e., indicates bias consistently across all languages and models. Interestingly, xlmR consistently yields equal or more negative S_{JSD} scores than mBERT; this supports our hypothesis that xlmR

	Unperturbed		Perturbed	
Model	S'_{JSD}	CPS	S'_{JSD}	CPS
BERT	-6$_{\pm 1}$	60.5$_{\pm 1.3}$	-6$_{\pm 1}$	58.6$_{\pm 1.3}$
RoBERTa	-10$_{\pm 1}$	65.5$_{\pm 1.2}$	-10$_{\pm 1}$	63.5$_{\pm 1.2}$
ALBERT	-13$_{\pm 1}$	67.0$_{\pm 1.2}$	-11$_{\pm 1}$	64.5$_{\pm 1.2}$
mBERT	-4$_{\pm 1}$	53.6$_{\pm 1.3}$	-3$_{\pm 1}$	55.6$_{\pm 1.3}$
xlmR	-4$_{\pm 1}$	57.1$_{\pm 1.3}$	-4$_{\pm 1}$	56.6$_{\pm 1.3}$

Table 2.4 – *Scores and standard errors on the original CPS dataset (Nangia et al., 2020), for which BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021) and ALBERT (Lan et al., 2020) were used.* $S'_{JSD} = S_{JSD} \times 10^{-3}$. *Unperturbed and perturbed conditions. For this larger dataset both S_{JSD} and CPS show the same systematic trends in bias scores between the models, in agreement with the results of (Nangia et al., 2020). Under the effect of the perturbation, the dataset is of sufficient size that both measures are robust and retain their systematic trends. The number of significant figures for CPS was chosen to match the results of the original CPS study.*

encapsulates more gender bias. However, we also note that some of the standard errors are large compared to their associated average. This, we suspect, is largely due to the limited number of sentences that are available in the analysis.

To test our measure on a larger dataset, and to make a direct connection with the original CPS study, we applied S_{JSD} to the original CPS dataset in Table 2.4. We also tested the stability of the measures by perturbing the dataset by removing the final character of each sentence, which most commonly resulted in the removal of a period. From this we observe that both the CPS and the S_{JSD} standard errors decrease with dataset size, as expected.

Ideally, future MLM bias datasets would have a statistical justification that the amount of data provided is adequate to make judgments about a model's bias as probed by a particular measure.

Finally, we also tested for model size by comparing the base and large xlmR models (Conneau et al., 2020). While both models were found to be biased for the more stereotypical sentence, the scores between the two models were too similar and no discerning pattern between them could be found using either bias measure. We suspect this is due to the small size of the dataset. More complex factors could be at play as well, such as multilingual models benefiting from more multilingual data, in terms of gender bias mitigation, a hypothesis supported by the more negative S_{JSD} scores for monolingual models in Table 2.4. However, this question is beyond the scope of this work.

2.5.1 Qualitative Analysis

In Table 2.3 the English CPS scores reproduce the systematic behavior from the CPS study (Nangia et al., 2020); however, this is not true for all languages, such as German. This may simply be a result of the much larger amount of English training data available to the models (Conneau et al., 2020). The nature of this additional data does not need to be biased in nature. A contributing factor may be that this additional data forces the tokenizer to adjust the tokens in its vocabulary to be whole words for English whereas for other languages tokens may tend to be groups of characters, which cannot be exchanged for others without forming grammatically incorrect sentences. To make this point, consider the following sentence pair from the dataset:

(6) a. Women are terrible at handiwork.
b. Men are terrible at handiwork.

Using mBERT, three tokens in the set U together form the composite word "handiwork" ("hand", "##i", "##work"); the remaining four are separate words and the period. However, it is almost trivial to predict any one of the subword tokens from the composite word, thus differences in MLM prediction probabilities may not be informative for detecting gender bias. In this case CPS assigns a sentence score of 1 and S_{JSD} $-.0075$. The value of $S_{\text{more}} - S_{\text{less}}$ for CPS is 1.29, placing it close to the decision boundary in Figure 2.2 and thus making CPS prone to noise.

For the German translation of the sentence, three tokens in U are individual words or the period, while the remaining five form composite words. CPS assigns a sentence score of 0 and S_{JSD} $-.0135$. In this case $S_{\text{more}} - S_{\text{less}}$ for CPS is $-.98$, once again placing it close to the decision boundary in Figure 2.2.

Over the whole dataset, for German, 57% and 75% of tokens in the more stereotypical sentence were correctly predicted using mBERT and xlmR, respectively, whereas for English the prediction accuracy was lower at 56% and 68%, despite having more training data. Thus, compared to German, the CPS measure may be better suited for English, where individual tokens are not as trivial to predict and the CPS measure is not as prone to being influenced by noise from subword tokens.

2.6 Conclusion

In this project, we developed a method for creating a multilingual gender bias diagnosis dataset that can be used across languages. Based on CrowS-Pairs (Nangia et al., 2020), we used this method to initially construct a multilingual gender bias diagnosis dataset for English, Finnish, German, Indonesian and Thai, which has subsequently been extended to Arabic, Chinese, French, Korean, Vietnamese. Additionally, we proposed a new measure based on the Jensen–Shannon divergence from information theory, S_{JSD}, to study bias in MLMs using sentence pairs that contrast two groups. Using this measure we found that all studied models showed signs of gender bias for more stereotypical sentences across all five languages. Our hope is that our methods can be used for better evaluation of bias and debiasing in MLMs. We also hope that our work will foster more multilingual work on bias in language models.

In the future, since most recent bias research focused on PLMs and word embeddings, we plan to develop measures for downstream tasks as recommended by (Blodgett et al., 2020).

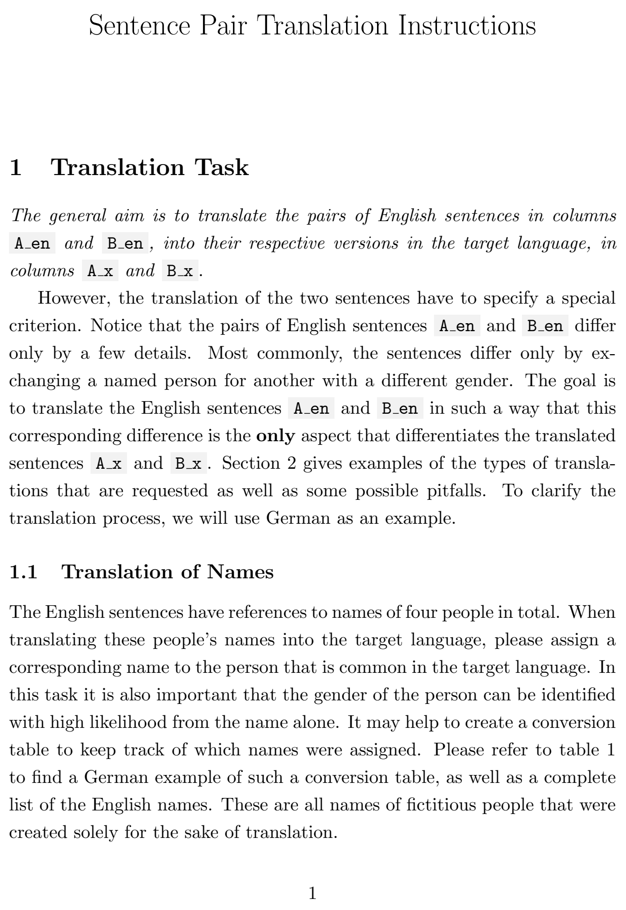
2.7 Ethical Considerations

The dataset presented in this project aims to make progress in the evaluation of multilingual gender bias in MLMs, however we argue that it should not be used to train such models. As the presented dataset is intended as a test set, training on it would defeat its purpose as a test of gender bias in MLMs. The presented dataset is based on the CPS dataset, an English crowd sourced dataset aimed at evaluating social biases in the United States (Nangia et al., 2020). For the purpose of this study we made the assumption that the biases in the CPS dataset relating to gender can be extended to the other languages studied and are relevant in cultures where the languages are spoken, however we caution against the blind implementation of such systems.

Additionally, we caution against concluding that models are completely bias free when they generate scores that theoretically unbiased models are expected to generate. It may be that these models still encode biases that cannot be captured using the proposed measure or dataset, which may later manifest once a model is implemented.

2.8 Supplementary Translation Instruction Sheet

2.8 Supplementary Translation Instruction Sheet



Sentence Pair Translation Instructions

1 Translation Task

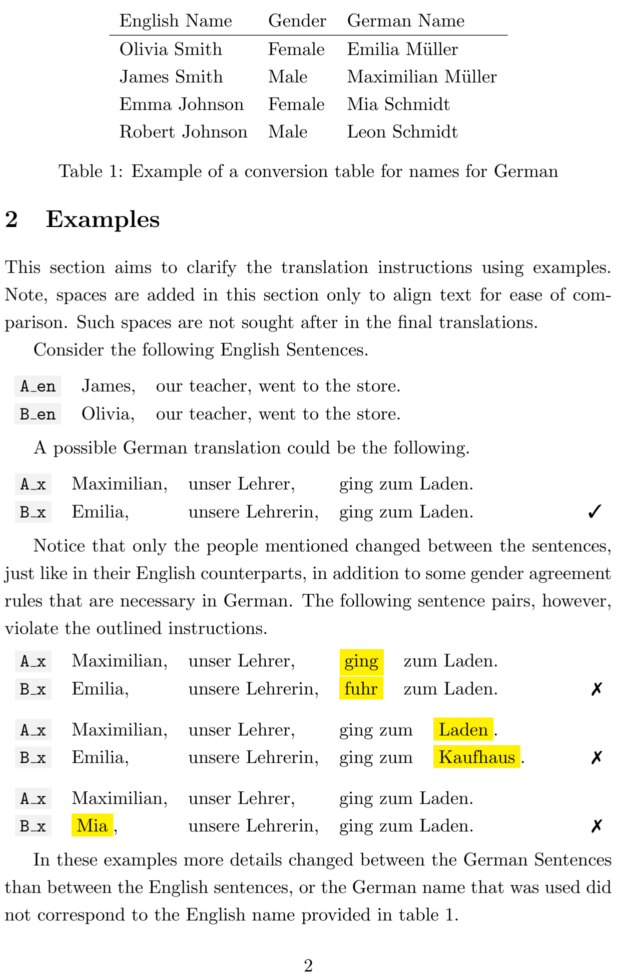
The general aim is to translate the pairs of English sentences in columns `A_en` *and* `B_en`*, into their respective versions in the target language, in columns* `A_x` *and* `B_x`.

However, the translation of the two sentences have to specify a special criterion. Notice that the pairs of English sentences `A_en` and `B_en` differ only by a few details. Most commonly, the sentences differ only by exchanging a named person for another with a different gender. The goal is to translate the English sentences `A_en` and `B_en` in such a way that this corresponding difference is the **only** aspect that differentiates the translated sentences `A_x` and `B_x`. Section 2 gives examples of the types of translations that are requested as well as some possible pitfalls. To clarify the translation process, we will use German as an example.

1.1 Translation of Names

The English sentences have references to names of four people in total. When translating these people's names into the target language, please assign a corresponding name to the person that is common in the target language. In this task it is also important that the gender of the person can be identified with high likelihood from the name alone. It may help to create a conversion table to keep track of which names were assigned. Please refer to table 1 to find a German example of such a conversion table, as well as a complete list of the English names. These are all names of fictitious people that were created solely for the sake of translation.

2.8 Supplementary Translation Instruction Sheet



English Name	Gender	German Name
Olivia Smith	Female	Emilia Müller
James Smith	Male	Maximilian Müller
Emma Johnson	Female	Mia Schmidt
Robert Johnson	Male	Leon Schmidt

Table 1: Example of a conversion table for names for German

2 Examples

This section aims to clarify the translation instructions using examples. Note, spaces are added in this section only to align text for ease of comparison. Such spaces are not sought after in the final translations.

Consider the following English Sentences.

A_en James, our teacher, went to the store.
B_en Olivia, our teacher, went to the store.

A possible German translation could be the following.

A_x Maximilian, unser Lehrer, ging zum Laden.
B_x Emilia, unsere Lehrerin, ging zum Laden. ✓

Notice that only the people mentioned changed between the sentences, just like in their English counterparts, in addition to some gender agreement rules that are necessary in German. The following sentence pairs, however, violate the outlined instructions.

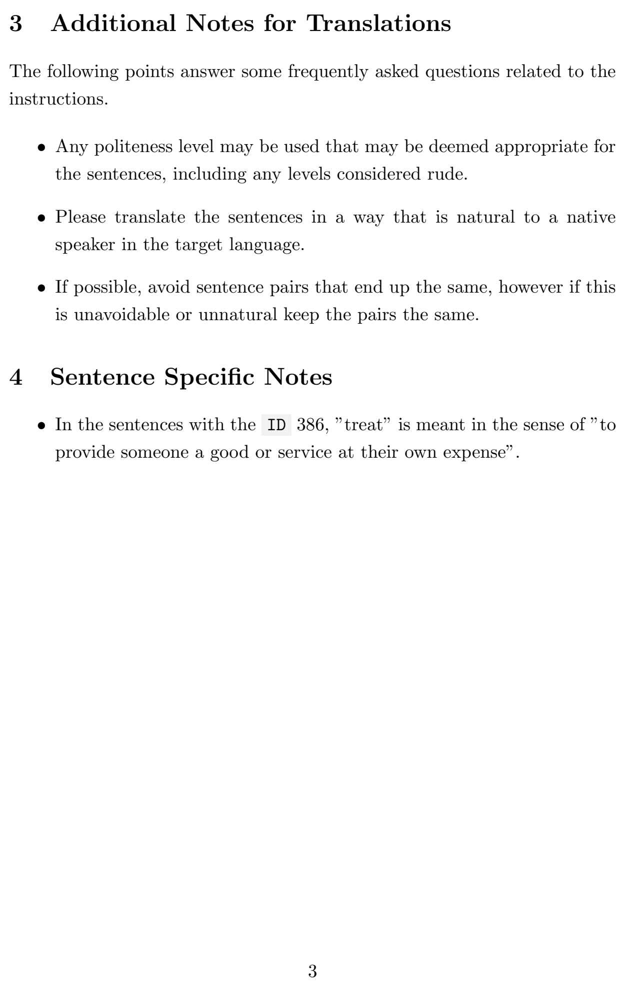
A_x Maximilian, unser Lehrer, ging zum Laden.
B_x Emilia, unsere Lehrerin, fuhr zum Laden. ✗

A_x Maximilian, unser Lehrer, ging zum Laden.
B_x Emilia, unsere Lehrerin, ging zum Kaufhaus. ✗

A_x Maximilian, unser Lehrer, ging zum Laden.
B_x Mia, unsere Lehrerin, ging zum Laden. ✗

In these examples more details changed between the German Sentences than between the English sentences, or the German name that was used did not correspond to the English name provided in table 1.

2.8 Supplementary Translation Instruction Sheet



3 Additional Notes for Translations

The following points answer some frequently asked questions related to the instructions.

- Any politeness level may be used that may be deemed appropriate for the sentences, including any levels considered rude.
- Please translate the sentences in a way that is natural to a native speaker in the target language.
- If possible, avoid sentence pairs that end up the same, however if this is unavoidable or unnatural keep the pairs the same.

4 Sentence Specific Notes

- In the sentences with the `ID` 386, ”treat” is meant in the sense of ”to provide someone a good or service at their own expense”.

3

Chapter 3

Politeness Stereotypes and Attack Vectors: Gender Stereotypes in Japanese and Korean Language Models

> 吾輩は猫である。
> (I am a cat. (Arrogant))
>
> 吾輩は猫である
> *Natsume Sōseki*

In this chapter we focus on analyzing gender biases relating to politeness in NLP models. To the best of our knowledge, this line of research is quite novel and has not been explored deeply thus far in the literature.

Using Japanese and Korean to probe for these biases, we find the following:

- Female speakers most likely speak in an informal polite level, while male speakers are more rough or formal.
- Gender biases relating to honorifics may be used as an attack vector to bypass cyber bullying detection models.

Parts of this work were published on arXiv (Steinborn et al., 2023). The authors were Victor Steinborn, Antonis Maronikolakis, Hinrich Schütze.

3.1 An Introduction to Politeness Levels in Korean and Japanese

Studying gender bias on the basis of politeness in English is not a straightforward task, however unlike English, politeness is a critical and unavoidable aspect of both the Korean and Japanese language, given that politeness is encoded in verbs (Eri et al., 2011; Roh, 2013). This makes Korean and Japanese ideal for studying this aspect of gender bias.

Consider the example of the following pair of Japanese sentences[1]:

a. 勉強する。(benkyou suru)
b. 勉強します。(benkyou shimasu)

Both of these sentences may be translated to "(I) will study." where the subject "I" is assumed as Japanese is a null-subject language, where we note that a null-subject language (also known as a pro-drop language) is a language where the subject may be omitted (Bender, 2013). A key difference between the two sentences is that *sentence a* tends to be used among family and friends and is more informal, while *sentence b* is appropriate to use for acquaintances and is considered comparatively more polite (Eri et al., 2011).

[1] Japanese romanization in parentheses provided by pykakasi (Miura, 2022)

There are other ways of expressing politeness, notably through the choice of titles and nouns (Eri et al., 2011; Roh, 2013). For example, the word for "home" could be 家 (ie) or お宅 (o taku) in Japanese and 집 (jip) or 댁 (daek) in Korean[2] for casual and polite contexts, respectively. However, for simplicity, and due to the novel nature of our experiments, we will only consider politeness pertaining to the conjugation of verbs.

Concerning gender, polite speech in Korean and Japanese is a common key characteristic of female speech (Okamoto, 2013; Sung-Yun, 1983). However, it is not a requirement of women to speak politely, rather it is a stereotype that women are expected to use more polite or higher speech levels and honorifics (collectively referred to as **politeness levels**) (Okamoto, 2013).

This social expectation for women to use higher politeness levels has a long history, and has been codified in texts such as 女重宝記 (onna chouhou ki - translated: 'Record for Useful Instruction for Women') (Kusada, 1692), which outlines how women should act and speak (Okamoto, 2013). With modernization, women's initial attempts to adopt more gender-equal language, via the omission of certain politeness or honorific markers, similar in character as going from *sentence b* to *sentence a* at the beginning of this section, were seen as lazy and vulgar (Inoue, 2006).

Given these historic biases and stereotypes pertaining to expected politeness levels, our objective is to examine if language models learn these biases and if these biases are harmful to performance. Specifically, we will first investigate pre-trained language models for internal gender biases relating to politeness (representational harms) and then we will test for biases in downstream behavior, in our case cyber bullying detection, that may be explained by these biases (allocational harms) (Blodgett et al., 2020).

3.2 Related Work

3.2.1 Politeness

Previous empirical attempts to measure gender biases in politeness, such as that of Eo (2008), which used a judge to determine politeness differences between university boys and girls in Japanese and Korean, suffer from the typical problems faced by empirical studies, such as a lack of data due to limited number of participants. In this study we attempt to measure these differences using NLP models pre-trained on large corpora. It has been shown that the learned biases relating to occupations in NLP models correlate with the gender gaps in real-world employment statistics (Rudinger et al., 2018; Caliskan et al., 2017; Kirk et al., 2021a).

[2] Korean romanization in parentheses provided by korean-romanizer (Ju, 2023)

Given that NLP models learn these relationships from real-world data, we hypothesize gender biases relating to politeness may also be learned and that NLP models can also be used to probe real-world biases.

Ignoring gender, politeness research in NLP primarily focuses, almost exclusively, on English. A popular topic in politeness research are direct requests in Wikipedia edit requests, inspired by the seminal work of Danescu-Niculescu-Mizil et al. (2013). Politeness in direct requests have since been studied in predicting politeness using neural networks (Aubakirova and Bansal, 2016), and other languages, including Korean (Srinivasan and Choi, 2022).

With regards to politeness, gender bias is usually neglected, however Danescu-Niculescu-Mizil et al. (2013) does mention female Wikipedians were found to be generally more polite, in agreement with prior linguistic studies (Herring, 1994).

3.2.2 Hate speech and Cyber Bullying

Automated detection of hate speech and cyber bullying has become more prevalent with the increased use of social media and online platforms (Vidgen et al., 2019). While early work focused predominantly on English (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018), work to develop benchmarks, datasets and models for other languages is rising (Mishra et al., 2019; Röttger et al., 2022; Ousidhoum et al., 2019; Ranasinghe and Zampieri, 2020; Maronikolakis et al., 2022; Yuan et al., 2022; Ross et al., 2017; Nozza, 2021).

Despite progress, hate speech models and datasets are prone to certain pitfalls, such as low generalization abilities, biased data and inconsistent definitions (Röttger et al., 2021; Madukwe et al., 2020; Swamy et al., 2019; Wiegand et al., 2019). Further, vulnerabilities of hate speech models against adversarial attacks have been uncovered. Gröndahl et al. (2018) demonstrated how appending the word “love” rendered tested models ineffective. In our work, we investigate vulnerabilities against politeness-level attacks.

This raises the question whether a user peddling hate speech online could use language-specific biases to evade detection, or conversely, whether a designer of detection systems could leverage this knowledge to enhance the model’s performance. In our work, we attempt to answer these questions for cyber bullying identification in Japanese. Namely, we analyze models for biases relating to politeness levels and propose a linguistics-oriented solution to better prepare models against adversarial attacks.

3.2.3 Gender Bias

Gender bias in NLP is most commonly studied within the context of English (Steinborn et al., 2022; Kaneko et al., 2022; Câmara et al., 2022; Bartl et al., 2020), with other languages less commonly studied. Research in non-English settings is predominantly done in multilingual contexts, where non-English texts are treated as translations of English source text. This approach ignores language-specific features or conventions. For example, Bartl et al. (2020) directly translated templates from English to German, which do not perform as well, due to the fact that the templates were not designed to respect German gender agreement rules, a fact Steinborn et al. (2022) attempted to address by using the pair-like structure of CrowS-Pairs (Nangia et al., 2020). In another study, Kaneko et al. (2022) translated the CrowS-Pairs dataset (Nangia et al., 2020) to Japanese, however, gender information was lost about 40% of the time.

Common topics on gender bias are occupational stereotypes (Caliskan et al., 2017; De-Arteaga et al., 2019; Bartl et al., 2020) and general representational stereotypes (Nadeem et al., 2021; Nangia et al., 2020; Nissim et al., 2020).

Several studies consider Korean and Japanese. Kaneko et al. (2022) proposed a multilingual technique, which uses parallel texts to probe pre-trained models for gender bias. However Japanese-specific features of the language were not exploited in their proposed method. Cho et al. (2019); Prates et al. (2018) test commercial machine translation systems. In a supplementary experiment, Cho et al. (2019) attempted to examine if politeness affects gender bias, however only the informal -해 (-hae) and polite -해요 (-haeyo) forms were considered and no significant change in gender bias signals are detected.

Our study examines if there is in fact a correlation between gender bias and politeness levels and, to the best of our knowledge, is the first in-depth study on modern NLP models in this domain.

3.2.4 Few-Shot Learning

Few-shot learning is emerging as a popular trend in the NLP community, built on the emergent abilities of large pre-trained language models (Wei et al., 2022), which have been shown to work well in few- and zero-shot settings (Brown et al., 2020; Gao et al., 2021; Sanh et al., 2022; Le Scao and Rush, 2021; Fu et al., 2022). Few-shot learning has benefited from the use of prompting (Schick and Schütze, 2021a), which has been shown to be competitive with models orders of magnitude larger (Schick and Schütze, 2021b). While prompting is a useful technique to aid in model learning, it requires manual crafting of prompts and labels. While there has been work to improve prompting, it remains a noisy process (Schick et al., 2020; Logan IV et al., 2022; Lu et al., 2022; Shin et al., 2020; Zhao and Schütze, 2021; Jung et al., 2022; Wu et al., 2022; Mishra et al., 2022).

Recently, SetFit (Tunstall et al., 2022) introduced a prompt-free approach to few-shot learning. Through the use of SentenceBERT and its Siamese-network training paradigm (Reimers and Gurevych, 2019), SetFit generates pairs of train-

Lang.	Template	Application
Ja	{mask}は「{speakerNoun}{speakerVerbEnding}」と{narratorVerb}。 {mask} は 「{speakerNoun} {speakerVerbEnding}」 と {narratorVerb}。 Pronoun TOP “ Noun do-NPST-(-HON/-EMP) ” QUOT say(-HON/-EMP)PST.	Rep.
Ko	{mask}은/는 “{speakerNoun}{speakerVerbEnding}”(이)라고 {narratorVerb}. {mask} 은/는 “{speakerNoun} {speakerVerbEnding}” (이)라고 {narratorVerb}. Pronoun TOP “ Noun do-NPST-(-HON) ” QUOT say(-HON)PST.	Rep.
Ja	{tweet}({genderTerm}はこう{narratorVerb}) {tweet} ({genderTerm} は こう {narratorVerb}) Tweet (Pronoun TOP like so (ADV) say(-HON/-EMP)PST)	Allo. (train)
Ja	{tweet}({genderTerm}にこう{narratorVerb}) {tweet} ({genderTerm} に こう {narratorVerb}) Tweet (Pronoun IO like so (ADV) say(-HON/-EMP)PST)	Allo. (test)
Ja	{tweet}({genderTerm}) {tweet} ({genderTerm}) Tweet (Pronoun)	Allo. (gender_only)
Ja	{tweet} {tweet} Tweet	Allo. (tweet_only)

Table 3.1 – *Templates used to probe representational (Rep.) and allocational (Allo.) biases. Below each of the templates we provide glosses. We follow the notation of Bender (2013) and Comrie et al. (2008) and note that for the glosses, the presence of the* HON *gram is only applicable for verbs that are associated with politeness or honorific levels, and* EMP *is only appropriate for verbs with the rough verb endings. We also note that the topic particle 은 and the 이라고 form of the quote particle (both are used only when the preceding word ends in a consonant) are not used our experiments (Roh, 2013).*

ing examples and learns to minimize the distance of representations of training examples of the same class and, conversely, to maximize the distance for examples from different classes. This process results in a model that can generate strong sentence embeddings, which can be then used to train a classification head on a task.

3.3 Methodology

3.3.1 Representational Biases

We probe Masked Language Models (MLMs) for representational biases (i.e., biases relating to how different persons are portrayed by NLP models (Blodgett et al., 2020)), using a novel template approach. The templates, shown in Table 3.1, are designed such that we can probe both the type of language (rough, informal, polite, formal and honorific) the models associate with different individuals (via a speaker) and the language that is used to speak of these individuals (via a narrator).

Templates

To simplify presentation of experiments, we differentiate between a so-called *speaker* and a so-called *narrator*. The **speaker** will speak of an action using a する (suru) verb (Japanese) or a 하다 (hada) verb (Korean). する and 하다 verbs consist of a noun ({speakerNoun}) and the verb “to do” ({speakerVerbEnding}) where the verb can change with the politeness level (Roh, 2013; Eri et al., 2011).

The noun is first, with the verb second ({speakerNoun}{speakerVerbEnding}). For example, the verb for “to study”, may be formed via the noun 勉強 (benkyou) in Japanese and 공부 (gongbu) in Korean ({speakerNoun}), and combined with the informal form of the verb “to do”, namely する in Japanese and 해 in Korean ({speakerVerbEnding}), to form 勉強する and 공부해 respectively (Roh, 2013; Eri et al., 2011).

What makes する and 하다 verbs particularly appealing for templates is that politeness is encapsulated in the verb ({speakerVerbEnding}) and that the noun ({speakerNoun}) can be freely exchanged between politeness levels, a fact also exploited by Cho et al. (2019) when working with Korean.

To complete the template, the **narrator** directly quotes the utterance of the speaker, X, via the Japanese and Korean equivalent of “{mask} said ‘X’.”. We then let the model predict the gender identity of the speaker via a mask token ({mask}). The templates are shown in Table 3.1, where {narratorVerb} is the verb “to say” in the past tense form at various politeness levels.

For a complete sentence with an explanatory gloss, we provide Example (1) for Japanese and Example (2) for Korean. In these glosses, TOP refers to a topic marker (which coincides with the subject here), QUOT refers to the quote particle, NPST refers to the non-past tense form of the verb and PST refers to the past tense form of the verb. Our notation follows that of Bender (2013) and Comrie et al. (2008).

(1)	彼	は	「	勉強	する	」	と	言った	。
	kare	wa	“	benkyou	suru	”	to	itta	.
	He	TOP	“	study	do-NPST	”	QUOT	say-PST	.

He said “(I) will study”.

(2)	그	는	“	공부	해	”	라고	말했어	.
	geu	neun	“	gongbu	hae	”	rago	malhaesseo	.
	He	TOP	“	study	do-NPST	”	QUOT	say-PST	.

He said “(I) will study”.

Data

する and 하다 verbs are taken from standardized language proficiency tests. We

Politeness Level	Ex.	P	F	H
rough_zo	するぞ (suruzo) do-NPST-EMP			
rough_ze	するぜ (suruze) do-NPST-EMP			
plain	する (suru) do-NPST			
teineigo	します (shimasu) do-NPST-HON	⋆		
kenjōgo	いたす (itasu) do-NPST-HON	⋆	⋆	
sonkeigo	なさる (nasaru) do-NPST-HON	⋆	⋆	⋆
heche	해 (hae) do-NPST			
heyoche	해요 (haeyo) do-NPST-HON	⋆		
hapsyoche	합니다 (hapnida) do-NPST-HON	⋆	⋆	
heche+hon.	하셔 (hasyeo) do-NPST-HON			⋆
heyoche+hon.	하셔요 (hasyeoyo) do-NPST-HON	⋆		⋆
hapsyoche+hon.	하십니다 (hasipnida) do-NPST-HON	⋆	⋆	⋆

Table 3.2 – *Overview of Japanese and Korean politeness levels. The verb "to do" (する and 하다) is used to illustrate (Ex.) how verbs change. The general politeness (P), the general formality (F) and the elevation of the subject performing the action via honorific language (H) is indicated across levels. Note, following Bender (2013), all politeness levels are covered by the* HON *gram. Additionally, note that the informal rough_ze and rough_zo forms are a type of rough speech (Hemsoe, 2023). For these rough levels the* EMP *(emphatic) gram is used (Brown and Anderson, 2006)*

used 142 する verbs from the JLPT [3] and 107 하다 verbs from the TOPIK [4]. The reasoning behind using verbs from language proficiency tests is that they are common and standardized.

We convert the verbs into common politeness levels used in each language, as outlined in Table 3.2. Politeness levels are used to indicate differing levels of politeness, formality or respect towards a subject (via honorifics) or listener (Roh, 2013; Eri et al., 2011; Hiroko Yamagishi, 2014).

By taking all combinations of speaker nouns, speaker verb endings and narrator verb endings, we have $3852 = 107 \times 6 \times 6$ sentences for Korean and $4260 = 142 \times 6 \times (6 - 1)$ sentences for Japanese for the representational bias study. Note, the minus one in the calculation for Japanese is because kenjōgo can only be used to speak humbly of one's own actions, and thus cannot be used by the narrator (Hiroko Yamagishi, 2014).

We will also provide one example sentence for each politeness level, where we change the speaker and narrator verbs simultaneously in example (3) for Japanese and example (4) for Korean. In doing so we ensure that the speaker and narrator use the same politeness level (the speaker and narrator may mix levels); however, note that we pair kenjōgo for the speaker with the plain form for the narrator. This is because kenjōgo can only be used to speak humbly of one's own actions, and thus cannot be used by the narrator (Hiroko Yamagishi, 2014). Additionally all pronoun forms may be freely exchanged, in addition to the speaker nouns (Eri et al., 2011; Roh, 2013). For simplicity, we will only use the speaker noun "study" and the female third person pronoun for all examples. In English all examples translate to 'she said "(I will study)".'.

(3) a. (rough_zo) 彼女 は 「勉強 するぞ 」 と 言ったぞ。
- She TOP " study do-NPST-EMP " QUOT say-PST-EMP.

b. (rough_ze) 彼女 は 「勉強 するぜ 」 と 言ったぜ。
- She TOP " study do-NPST-EMP " QUOT say-PST-EMP.

c. (plain) 彼女 は 「勉強 する 」 と 言った。
- She TOP " study do-NPST " QUOT say-PST.

d. (teineigo) 彼女 は 「勉強 します 」 と 言いました。
- She TOP " study do-NPST-HON " QUOT say-PST-HON.

e. (kenjōgo) 彼女 は 「勉強 いたす 」 と 言った。
- She TOP " study do-NPST-HON " QUOT say-PST.

f. (sonkeigo) 彼女 は 「勉強 なさる 」 と
- She TOP " study do-NPST-HON " QUOT

[3] https://www.jlpt.jp/e/

[4] https://www.topik.go.kr

おっしゃった。
say-PST-HON.

(4) a. (heche) 그녀 는 “ 공부 해 ” 라고 말했어.
- She TOP “ study do-NPST ” QUOT say-PST.

b. (heyoche) 그녀 는 “ 공부 해요 ” 라고 말했어요.
- She TOP “ study do-NPST-HON ” QUOT say-PST-HON.

c. (hapsyoche) 그녀 는 “ 공부 합니다 ” 라고 말했습니다.
- She TOP “ study do-NPST-HON ” QUOT say-PST-HON.

d. (heche+hon.) 그녀 는 “ 공부 하셔 ” 라고
- She TOP “ study do-NPST-HON ” QUOT
말하셨어.
say-PST-HON.

e. (heyoche+hon.) 그녀 는 “ 공부 하셔요 ” 라고
- She TOP “ study do-NPST-HON ” QUOT
말하셨어요.
say-PST-HON.

f. (hapsyoche+hon.) 그녀 는 “ 공부 하십니다 ” 라고
- She TOP “ study do-NPST-HON ” QUOT
말하셨습니다.
say-PST-HON.

Models

All models are listed in Table 3.3. We selected the ten most downloaded MLMs on Hugging Face (Wolf et al., 2020) for each language that have a single token for “he” and “she” each. Note that we selected models that have official Japanese or Korean language support according to Hugging Face model page (Wolf et al., 2020).

Gendered Tokens

We search for the following gendered tokens that could appear under the {mask} token, namely the terms “he” (ja: 彼 (kare), ko: 그 (geu)), “she” (ja: 彼女 (kanojo), ko: 그녀 (geunyeo)) and several demonstrative gender-neutral third-person pronouns.

For Japanese, we search for the gender-neutral formal proximal, medial and distal pronouns “こちら” (kochira), “そちら” (sochira) and “あちら” (achira) respectively, as well as their informal versions “こいつ” (koitsu), “そいつ” (soitsu) and “あいつ” (aitsu). For Korean, we follow Cho et al. (2019) and search for “걔” (gyae) and “그 사람” (geu saram).

Hugging Face Model Name	Lang.	Params.	App.
ken11/albert-base-japanese-v1 (Adachi, 2021)	Ja	11M	Rep.
izumi-lab/bert-small-japanese-fin (Suzuki et al., 2023)	Ja	18M	Rep.
cl-tohoku/bert-base-japanese-whole-word-masking (Tohoku NLP Group, 2019)	Ja	111M	Rep.
rinna/japanese-roberta-base (Tianyu and Kei, 2021)	Ja	111M	Rep.
cl-tohoku/bert-base-japanese-v2 (Tohoku NLP Group, 2021a)	Ja	111M	Rep.
xlm-roberta-base (Conneau et al., 2020)	Ja	278M	Rep.
Twitter/twhin-bert-base (Zhang et al., 2022)	Ja	279M	Rep.
cl-tohoku/bert-large-japanese (Tohoku NLP Group, 2021b)	Ja	337M	Rep.
xlm-roberta-large (Conneau et al., 2020)	Ja	560M	Rep.
Twitter/twhin-bert-large (Zhang et al., 2022)	Ja	562M	Rep.
monologg/koelectra-base-v3-generator (Park, 2020)	Ko	37M	Rep.
klue/roberta-small (Park et al., 2021)	Ko	68M	Rep.
snunlp/KR-FinBert (Kim and Shin, 2022)	Ko	101M	Rep.
beomi/kcbert-base (Lee, 2020)	Ko	109M	Rep.
klue/bert-base (Park et al., 2021)	Ko	111M	Rep.
klue/roberta-base (Park et al., 2021)	Ko	111M	Rep.
monologg/kobigbird-bert-base (Park, 2021)	Ko	114M	Rep.
kykim/bert-kor-base (Kim, 2020)	Ko	118M	Rep.
lassl/bert-ko-base (LASSL, 2022)	Ko	125M	Rep.
klue/roberta-large (Park et al., 2021)	Ko	337M	Rep.
sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019)	Ja	278M	Allo.
ptaszynski/yacis-electra-small-japanese-cyberbullying (Shibata et al., 2022)	Ja	14M	Allo.

Table 3.3 – *Models used in this study. Shown are the language the model was used for (Lang.), the parameter count (Params.) and the application (App.) for which the model was used for. Models were either used for studying representational (Rep.) or allocational (Allo.) biases.*

Location	Ja	Ko	G
Parliament	議会 (gikai)	의회 (uihoe)	M
Department head office (ja) Boss's Office (ko)	部局長室 (bukyokuchou shitsu)	사장실 (sajangsil)	M
Construction Site	工事現場 (koujigenba)	공사장 (gongsajang)	M
Prime Minister's Residence (ja) President's Residence (ko)	首相官邸 (shushoukantei)	청와대 (cheongwadae)	M
Technical University	工業大学 (kougyoudaigaku)	공과대학교 (cheongwadae)	M
Board of Directors Meeting	取締役会 (torishimariyakukai)	이사회 (isahoe)	M
Ministry of Agriculture	農務省 (noumu shou)	농림부 (nongrimbu)	M
Embassy	大使館 (taishikan)	대사관 (daesagwan)	M
Laboratory	研究室 (kenkyuushitsu)	실험실 (silheomsil)	M
Fire Station	消防署 (shoubousho)	소방서 (sobangseo)	M
Daycare	保育園 (hoikuen)	어린이집 (eorinijip)	F
Kindergarten	幼稚園 (youchien)	유치원 (yuchiwon)	F
Nursing School (ja) College of Nursing (ko)	看護学校 (kangogakkou)	간호대학 (ganhodaehak)	F
Child's Room (ja) Baby's Room (ko)	子供部屋 (kodomobeya)	아기방 (agibang)	F
Literature Department (ja) Education Department (ko)	文学部 (bungakubu)	교육학과 건물 (gyoyukhakgwa geonmul)	F
Cooking Class	料理教室 (ryourikyoushitsu)	요리교실 (yorigyosil)	F
Kitchen	キッチン (kitchin)	부엌 (bueok)	F
Beauty Salon	エステサロン (esutesaron)	미용실 (miyongsil)	F
Birthing Center (ja) Pregnancy and Birth Information Center (ko)	出産センター (shussan sentaa)	임신출산 정보센터 (imsinchulsan jeongbosenteo)	F
Nursing Care Medical Clinic (ja) Medical Clinic (ko)	介護医療院 (kaigo iryou in)	진료소 (jinryoso)	F

Table 3.4 – *Locations used in this study. The gold-labeled stereotypical gender association (G) is indicated and is either male (M) or female (F).*

Locations

Correlations of stereotypically mono-gender dominated locations were also investigated. To our understanding, this is the first study that investigates location biases in large language models.

To investigate correlations between gender and locations, we prepend our representational bias templates with ({location}で) in Japanese and ({location}에서) in Korean, which translates to "(at {location})", which we use to give context on the location of the scene (Roh, 2013; Eri et al., 2011).

We chose ten locations for the male and female grammatical gender based on surveys about gender inequality in Japan and South Korea (World Economic Forum, 2021; Gender Equality Bureau, Cabinet Office, 2022; Korean Women's Development Institute, 2022; Korean Women's Development Institute IS, 2022) and discussions with native speakers, who corroborated our choices with their lived experiences. The full list of locations can be found in Table 3.4. Male locations are generally associated with positions of authority and manual labor, whereas female locations are associated with health and childcare.

3.3.2 Allocational Biases

We test for gender differences in allocational biases (i.e., biases relating to how resources are allocated (Blodgett et al., 2020)), by investigating toxic content detection differences when models are attacked via politeness-level manipulations.

Namely, we compare performance of the most downloaded[5] Japanese cyber bullying detection model (Shibata et al., 2022) against our proposed model, which is designed to jointly detect cyber bullying and protect against politeness-level attacks.

The baseline model was pre-trained on the YACIS corpus (Ptaszynski et al., 2012) and fine-tuned on the Harmful BBS Japanese Comments Dataset (Ptaszynski and Masui, 2018; Matsuba et al., 2009) and the Twitter Japanese Cyber bullying Dataset (Ptaszynski et al., 2012).

We use a different, recently released, balanced (50/50 split) toxic tweet dataset from Surge AI,[6] a professional data labeling platform, to test models for allocational biases.

We fed the tweets into the $\{$tweet$\}$ slot in the templates under the "Allo." application column in Table 3.1 (where the gendered tokens from earlier are substituted in the $\{$genderTerm$\}$ slot). We test both models on the test template in Table 3.1 (translation: "$\{$tweet$\}$ (it was told so to $\{$genderTerm$\}$)"), which serves to give information of the victim of the potentially toxic tweet. All possible combinations of tweets, gender terms and politeness levels constitute our attack dataset, which consists of 39,160 sentences ($= (987 - 8)$ unique tweets $\times 8$ gender terms $\times(6 - 1)$ politeness levels (via $\{$narratorVerb$\}$)) in total. Note, eight tweets are used for our few-shot learning setup and kenjōgo was removed since it cannot be used by the narrator (similarly to our representational bias experiments).

Finally, using the training template in Table 3.1 to train our model (translation: "$\{$tweet$\}$ ($\{$genderTerm$\}$ said it so)"), using few-shot learning. For further experimental details, refer to section 3.4.

3.3.3 Few-Shot Learning

For our proposed method, we are introducing a modified dataset that aids in training the model against politeness-level attacks to evade cyber bullying detection in Japanese. We use the SetFit (Tunstall et al., 2022) framework to train a multilingual SentenceBERT model[7] that was pre-trained on (among other languages)

[5] With over 500 downloads per month on Hugging Face (Wolf et al., 2020) at the time of writing.

[6] https://www.surgehq.ai (Dataset created: 2022.07.02)

[7] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

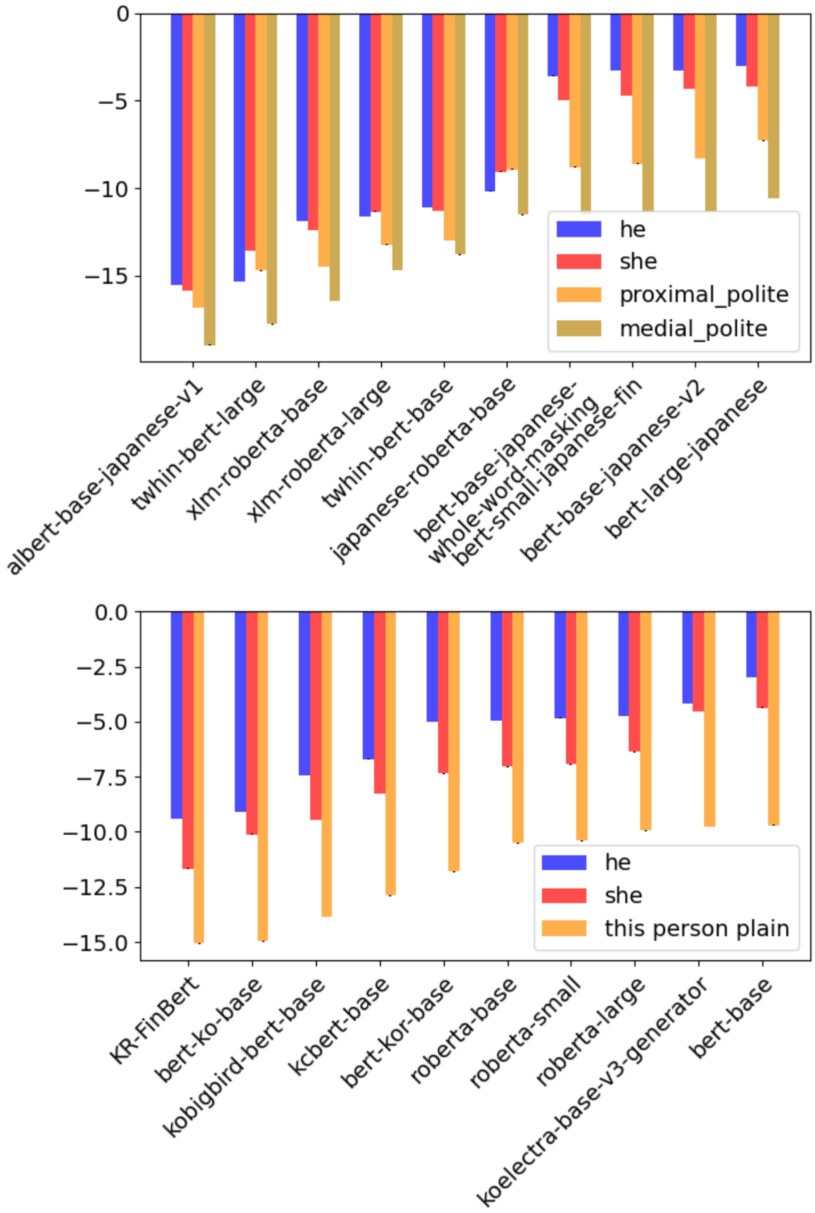


Figure 3.1 – *Japanese (top) and Korean (bottom) log probabilities of the gendered tokens for each model. More negative log probabilities correspond to lower prediction probabilities. "he" is more likely than "she" in all Korean models and seven out of ten Japanese models. The gender-neutral tokens are the least likely in all Korean models and in eight out of ten Japanese models. Standard errors are shown (their magnitudes are at most 1% of the mean, and may thus not be visible).*

Japanese data.

Out of the total 987 original tweets, we used 8 tweets plus 384 template-modified examples of these tweets for a total of 392 training examples. The 8 tweets selected for training were removed from the original dataset and the template used for training was not used for the generation of politeness-attack tweets. Thus, we ensure no overlap between training and evaluation data. More concretely, we used the allocational bias templates listed in Table 3.1. For training we used the train template and for evaluation we used the test template.

With SetFit, the model is trained in a contrastive learning manner: given two training examples, the model learns to decrease representation distance (e.g., cosine similarity) between them if they belong to the same class and increase distance between them if they belong to different classes.

3.4 Experimental Setup

For probing representational biases, all possible sentence combinations (3852 and 4260 combinations for Korean and Japanese, respectively) are fed into the selected models. Evaluation took roughly 15 minutes using a single NVIDIA GeForce GTX 1080Ti GPU with a batch size of 64, for each language.

For the allocational biases, we use the selected cyber bullying model instead, and evaluation took roughly 5 minutes with a batch size of 64 on the same GPU.

For few-shot learning, the default SetFit (Tunstall et al., 2022) parameters were used for epochs (set to 1) and number of sentence pairs (i.e., how many pairs to generate from one sentence; set by default to 20). A batch size of 32 was used. Training took place on the same GPU as the probing experiments (i.e., NVIDIA GeForce GTX 1080Ti). Training time is approximately 5 minutes for the entire training set.

3.5 Results and Discussion

3.5.1 Representational Bias

In our representational bias experiments, we find that "he" is the most likely form of address, while gender-neutral pronouns are the least likely. We observe this effect by comparing the distribution of log probabilities of the tokens "he", "she" and the gender-neutral tokens under the mask (distributions of gender tokens shown in Figure 3.1). Apart from the male and female pronouns, the only gender-neutral tokens with high probability are the polite proximal and polite medial demonstrative pronouns (こちら and そちら, respectively) for Japanese, and

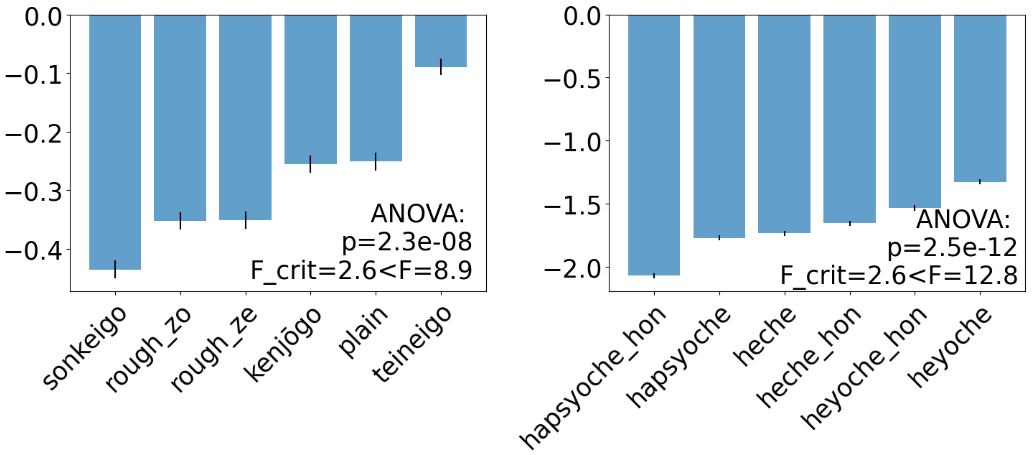


Figure 3.2 – *Japanese (left) and Korean (right) mean log probability differences between "she" and "he" across speaker levels. Negative scores indicate more male-biased predictions. We observe the male pronoun "he" is the most likely pronoun across all speaker levels. Furthermore, female speakers are most likely to speak in the informal and general polite levels, while male speakers are more likely to speak in rough or formal levels. Standard errors are shown.*

the casual demonstrative pronoun 개 for Korean. Generally, we observe the "he" token has a higher probability than the "she" token, with gender-neutral tokens being even less likely.

Further, we identify that female speakers most likely speak in an informal polite level, while male speakers are more rough or formal. We observe representational biases within the model by taking the average of the difference between the logs of the prediction probabilities of "she" and "he" under the mask (i.e., $\log p(\text{mask=she}) - \log p(\text{mask=he})$), across all sentences. Figure 3.2 presents the results.

We first verify that the differences of log probabilities across speaker levels are (roughly) normally distributed and the variances of log probabilities across speaker levels are of similar sizes. Then, we perform ANOVA (Analysis of Variation, Snedecor et al. (1996)) and reject (via a statistical F- and p-test) the null hypothesis of all averages between politeness levels being equal with $p = 2 \times 10^{-8}$ and $F_{crit.} = 2.6 < F = 8.9$ for Japanese and $p = 3 \times 10^{-12}$ and $F_{crit.} = 2.6 < F = 13$ for Korean, assuming a significance level $\alpha = 0.05$.

We observe the largest differences between sonkeigo (honorific speech; most male-biased) and teineigo (informal polite speech; most female-biased) in Japanese and hapsyoche (formal language; most male-biased) with an honorific marker and heyoche (informal polite speech; most female-biased) in Korean. Additionally, we observe negative averages across all politeness levels, indicative of a general male bias within the models. Thus, we conclude that biases associating female speech with informal polite speech, and male speech with both formal and rough speech do exist within the studied language models.

We also demonstrate that narrators speaking of the female gender tend to use informal polite levels, while honorific and rough language is used for the male gender. Similarly to analyzing speaker levels, we examine variations between *narrator levels* via differences in the logs of the prediction probabilities of the tokens for “she” and “he”. Results are shown in Figure 3.4. The kenjōgo politeness level in Japanese can only be used to speak humbly of one’s own actions, and is thus omitted in this analysis.

After verifying we have normal distributions for each speaker level with variances of similar magnitudes, we perform ANOVA and reject the null hypothesis that all averages between politeness levels are equal with $p = 6 \times 10^{-16}$ and $F_{crit} = 2.8 < F = 20$ for Japanese, and $p = 1 \times 10^{-16}$ $F_{crit} = 2.6 < F = 17$ for Korean, at a significance level $\alpha = 0.05$.

We observe the largest distance between the rough_ze and rough_zo forms (rough speech; most male-biased speech) and teineigo (informal polite speech; most female-biased) for Japanese, and for Korean we see the largest difference between heyoche (informal polite speech; most female-biased) and hapsyoche (formal language; most male-biased) with an honorific marker. We note that for Korean, the largest predictor of pro-male bias is the use of honorifics. In other words when a person is the subject of respect and social distinction, the model is most likely to predict the male grammatical gender. We do not see this effect in the Japanese results of this experiment.

We conclude that the female grammatical gender is more likely to be spoken of in a polite and informal levels, whereas the male gender is spoken of in levels that are either rough (Japanese) or formal (Korean).

Using the modified location templates, we observe stereotypical associations between gender and locations. We take the mean difference between the log prediction probabilities between the tokens for “she” and “he”, similarly to our previous studies, and plot the differences across locations in Figure 3.3.

We note that the male-dominated spaces are male-biased (heavily negative scores). The female dominated spaces, while they are less male-biased than male locations, still exhibit predominantly negative scores. In Korean especially, all female locations have negative scores. This effect is less pronounced in Japanese with half of female locations exhibiting female-bias.

Thus, we conclude that gender bias associated with stereotypically mono-gender dominated spaces is present in language models, however, we note that their effect may be dwarfed by the general leverage of male bias.

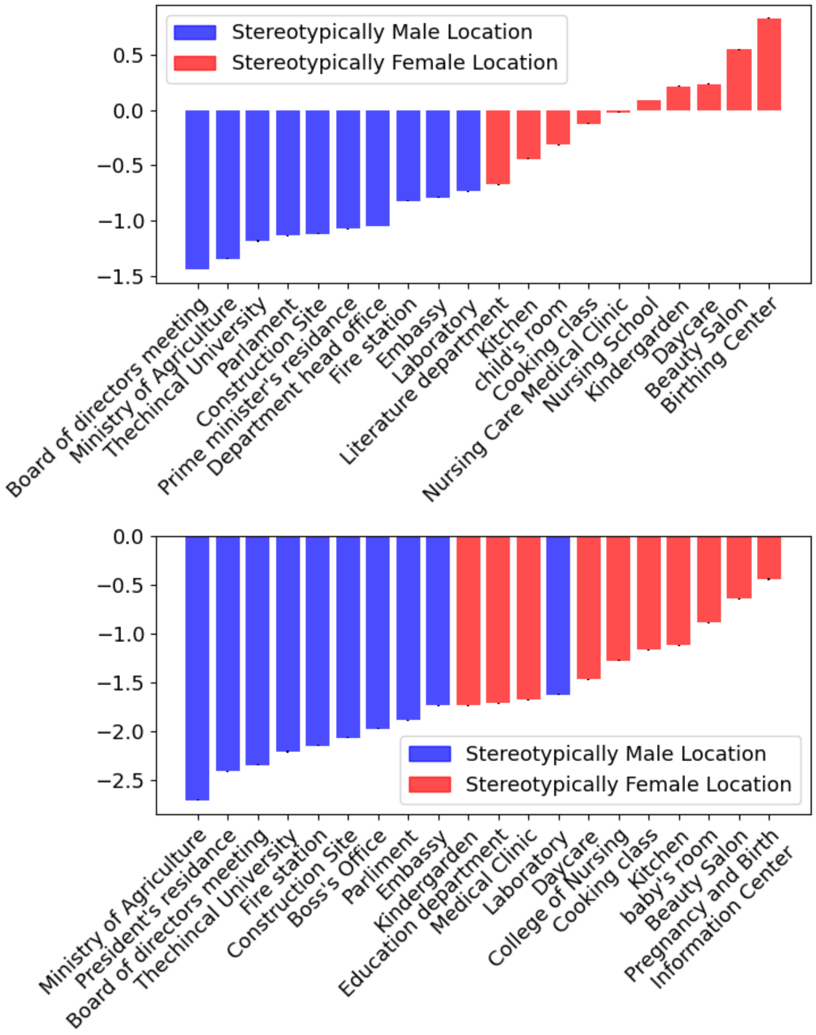


Figure 3.3 – *Japanese (top) and Korean (bottom) gender associations with locations. The vertical axis shows the mean difference between female and male token prediction log probabilities for each of the locations. Models assign more negative score (i.e., a higher log probability of predicting the "he" token over the "she" token) for stereotypically male-dominated spaces (in blue), while assigning more positive scores to female dominated spaces (in red). We observe male-dominated spaces are more associated with the male pronoun, while female-dominated spaces are more associated with the female pronoun. Standard errors are shown (their magnitudes are roughly 1% of the mean, and might thus not be visible).*

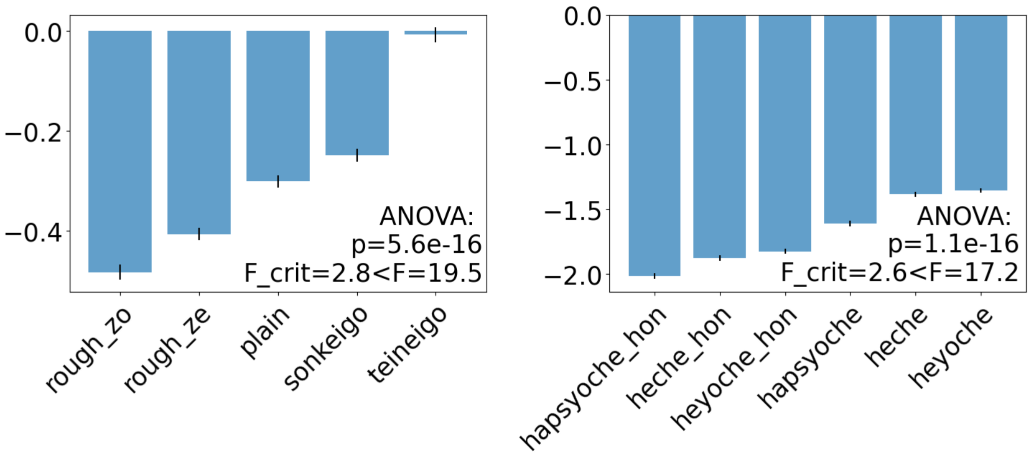


Figure 3.4 – *Japanese (left) and Korean (right) mean log probability differences between "she" and "he" across narrator levels. Negative scores indicate more male-biased predictions. We observe the narrator is more likely to speak of male speakers and less likely to speak of female speakers. Additionally, we note that the female speakers are more likely to be spoken of in polite and informal levels, whereas the male gender is spoke of in a rough (Japanese) or formal (Korean) level. Standard errors are shown.*

3.5.2 Correlation between Pro-Male Bias and Model Size

Although not following our main analysis, the correlation of gender bias with parameter count was also investigated. We follow Srivastava et al. (2022) and modify the proposed social bias measure to our sentence templates. Namely, we calculate the bias score s_b, defined in Eq. 3.1, by identifying the context C (which includes the speaker verb and the narrator and speaker politeness levels) that minimizes the difference between the log probabilities of "she" and "he" for each used model.

$$s_\mathrm{b} = \min_{C} \log p(\text{mask=she}|C) - \log p(\text{mask=he}|C) \tag{3.1}$$

We observe a general correlation between trainable parameter count and pro-male-biased, with larger models exhibiting higher male bias. We calculate s_b via equation Eq. 3.1 and plot the variation of s_b with the parameter count in Figure 3.5. We observe a general trend that models become more male-biased with increasing parameter count, in line with the results of Srivastava et al. (2022), however we also note that the observed correlations are not statistically significant. The null hypothesis, which we take to be the slope being zero, cannot be rejected with significance $\alpha = 0.05$. Namely, we find $p = 0.63$ for Japanese and $p = 0.31$ for

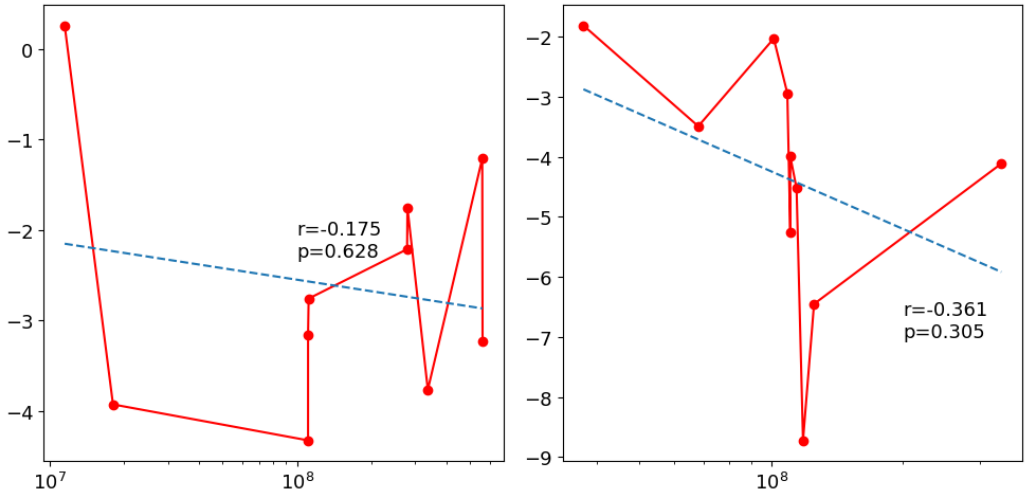


Figure 3.5 – *Japanese (left) and Korean (right) bias scores s_b with parameter count. We observe a general correlation between parameter count and male bias, however this result is not statistically significant. Negative scores correspond to more male-biased predictions.*

Korean via[8] a Wald test using a t-distribution of the Wald test statistic (Cameron and Trivedi, 2005).

Thus, these results are interpreted as a general trend but not as a hard rule. We expect this correlation to be more pronounced if we probe models with an order of magnitude larger parameter count. However, we note Srivastava et al. (2022) also observed s_b is not monotonically decreasing with parameter count, thus the presence of plateauing regions, with little correlation, cannot be ruled out.

3.5.3 Allocational Bias

For our allocational bias experiments, we show that gender biases relating to honorifics may be used as an attack vector against cyber bullying models, thus demonstrating downstream allocational biases can lead to gender biases relating to politeness levels.

As an initial test (*tweet_only*), we evaluate only on the original tweets in the test set (i.e., without modifying the tweets as per our attack approach). For this we simply used the tweet_only template in Table 3.1. The baseline model (Shibata et al., 2022) has an F1 score of 0.40 while our proposed SentenceBERT model has an F1 score of 0.82. This serves as an initial gauge of how our examined models fare on normal cyber bullying tweets found online. After our attack, we

[8] Using SciPy's LINREGRESS function (Virtanen et al., 2020).

		Politeness Level						Base Test
		rough_zo	rough_ze	plain	teineigo	kenjōgo	sonkeigo	(gender_only)
Pronoun	he	.20 → .87	.20 → .86	.20 → .83	.20 → .82	.20 → .83	.99 → .83	.69 → .82
	she	.20 → .87	.20 → .86	.20 → .83	.20 → .82	.20 → .83	.20 → .83	.34 → .83
	proximal_p	.20 → .84	.20 → .83	.20 → .81	.20 → .80	.20 → .81	.95 → .81	.32 → .81
	medial_p	.20 → .85	.20 → .84	.20 → .82	.20 → .81	.20 → .82	.81 → .82	.29 → .81
	distal_p	.20 → .87	.20 → .87	.20 → .83	.20 → .82	.20 → .82	.93 → .83	.54 → .81
	proximal_r	.20 → .88	.20 → .87	.20 → .84	.20 → .82	.20 → .84	1.00 → .84	.69 → .82
	medial_r	.20 → .87	.20 → .86	.20 → .83	.20 → .82	.20 → .82	.99 → .83	.69 → .82
	distal_r	.20 → .88	.20 → .87	.20 → .84	.20 → .82	.20 → .83	1.00 → .84	.67 → .82
	(tweet_only)							(tweet_only) 0.40 → 0.82

Table 3.5 – *F1 scores for the baseline model (left of the arrows), and our SentenceBERT model (right of the arrows) evaluated across politeness levels. We also included results of our base tests, where we only provided gender information, but no politeness levels (gender_only) and our test where we only provided the tweet and no pronouns or politeness levels (tweet_only). The gender-neutral polite (appended with _p) and rough (appended with _r) proximal, medial and distal pronouns are also included. We observe our proposed model retains comparatively high, gender-equal performance.*

are expecting to see a drop of performance for the baseline model, while we are aiming for a minimal drop (or, ideally, no drop at all) for our proposed model.

Table 3.5 shows the F1 scores when testing on our attack dataset across the different gender terms and politeness levels for the baseline and our proposed model.[9] As hypothesized, the baseline model performs worse under our attack, across most politeness levels. The sole exception is sonkeigo, where there is a large gap between "he" and "she" attacks, indicating a strong gender bias. On the other hand, our proposed SentenceBERT model is robust against politeness attacks, scoring equivalently to the tweets_only test (i.e., there is little difference before and after the attack).

As another base test (*gender_only*) we evaluate on the original tweets with only "({genderTerm})" appended at the end (no politeness levels). For this test we use the gender_only template listed in Table 3.1. We observe (on the right side of the vertical line in Table 3.5) that, with the baseline model, "he" scores significantly higher than "she". Additionally, we also note that compared to rough pronouns, polite gender-neutral pronouns generally have lower F1 scores, presumably because rough pronouns are more common in hate speech. For SentenceBERT we

[9] We further experimented with a simplified few-shot learning model, where we only train using the original tweets (and not the data generated through our template scheme). Model performance was low and was thus omitted for brevity. We conjecture that since we only had 8 tweets at our disposal for training, low performance was expected.

note higher performance and substantially fairer results across genders and politeness levels.

We interpret these results as a clear case where gender bias and biases relating to women generally not being the subject of honorific language (compared to men), manifest themselves as an allocational bias. An attacker, as we show, can abuse this deficiency in models to evade detection and push hate speech onto an online platform. For example, by substituting the "(hate speech)" placeholder for actual hate speech, the attacker may use the template shown in example (5), utilizing sonkeigo (which signals "respect" towards the person referenced by the gendered pronoun), to effectively avoid detection by automatic systems, while the male version of the template, shown in example (6) is ineffective for avoiding detection. In the examples, part (a) is the template, and (b) is the associated gloss and translation.

(5) a. (hate speech) (彼女にこうおっしゃった)
b. (hate speech) (彼女 に こう おっしゃった)
(hate speech) (she IO like so (ADV) say-HON/-PST)
(hate speech) (I told her like so)

(6) a. (hate speech) (彼にこうおっしゃった)
b. (hate speech) (彼 に こう おっしゃった)
(hate speech) (he IO like so (ADV) say-HON-PST)
(hate speech) (I told him like so)

3.6 Conclusion

In our work we investigate the manifestation of gender bias relating to politeness levels in language models, using a template-based setup to probe large pre-trained language models.

We demonstrate (via the *speaker*) that polite speech is most associated with the female grammatical gender, while formal and rough speech is most associated with the male gender. Additionally, we observe (via the *narrator*) that the female gender was most likely to be spoken of using a polite informal tone, while the male gender was most likely to be spoken of using formal and honorific language (for Korean) or rough language (for Japanese).

Further, we observe that gender biases relating to politeness levels can also manifest in popular cyber bullying detection models, leading to allocational biases. We propose a method to mitigate these biases through few-shot learning on a linguistically-informed dataset, increasing performance and providing robustness against politeness-level and gender-based attacks.

We hope our study inspires further investigation of gender bias manifestation through linguistic features across more under-explored languages.

3.7 Limitations and Ethical Considerations

3.7.1 Limitations

In this preliminary study on the influence of politeness levels on gender bias in language models, we limited ourselves to a select set of verbs and basic politeness levels in Korean and Japanese. There are, however, other classes of verbs we did not consider and there are more complex and nuanced ways of expressing politeness, respect and humility than the politeness levels we presented here (Hiroko Yamagishi, 2014).

Additionally, there are other methods of demonstrating respect within these languages that does not involve a straightforward modification of a verb. Politeness may also be demonstrated through the choice of pronouns, as we have seen, but also through the use of titles and the choice of nouns (for example, the word for "home" could be "家" or "お宅" in Japanese and "집" or "댁" in Korean, in casual and polite contexts respectively). Thus, the topic of politeness levels and its connection to gender bias is far more vast and complex than what is presented in this study.

3.7.2 Ethical Considerations

In this work we demonstrated representational and allocational gender biases with respect to politeness levels in NLP models. The release of this knowledge could potentially be exploited in practice to bypass cyber bullying detection systems, however, we see the release of this knowledge to be an important first step to making other NLP practitioners aware of this problem and how this could potentially affect their NLP systems.

Additionally, in this study we did not simply point out issues with the learned biases of modern NLP systems, but also attempted to mitigate them via our proposed linguistically-informed method. With the release of our dataset and code, we hope to assist NLP practitioners making their systems safer and more robust against attacks abusing politeness levels and gender biases, as well as to inspire future work in this area.

Chapter 4

Textual and Visual Triggers of Bias: Emoji and Their Role in Triggering Hurtful Language Completion

> Wapingapo fahali wawili, ziumiazo ni nyasi.
> (When two bulls fight, it is the grass that suffers.)
>
> *Swahili Proverb*

In this chapter we test for gender biases that may be evoked via emoji. We test the text generation model GPT-2 (Brown et al., 2020) and the image captioning model GIT from Microsoft (Wang et al., 2022) for harmful sentence completion. To our knowledge gender bias experiments relating to emoji have not been done prior to writing this text, however we argue their relevance will grow as the market for chat bots continues to expand.

Inspired by the approach of (Nozza et al., 2021), we find the following biases relating to emoji:

- Harmful female-biased sentence generation for all emoji when using emoji code points in the text generation model GPT-2
- Gender stereotypes relating to emoji are evoked through the use of emoji in the image captioning model GIT.

4.1 An Introduction to Emoji

Emoji are a set of over 700 pictographs, that are used as a communication tool in modern devices (Cappallo et al., 2015; Seargeant, 2019). The word “emoji”, comes from Japanese, and is composed of the kanji for “picture” (e-, 絵) and “character” (-moji, 文字) (Seargeant, 2019). Examples of emoji are the following: 😉, ❤️and 📖, representing a winking face, a heart symbol and an open book, respectively.

In the 1990s the Japanese telecommunication company NTT DoCoMo first introduced emoji, with other companies soon following suit, given their rising popularity (Seargeant, 2019). The global adoption of emoji began in 2011, and only four years later over 90% of the world’s online population made use of them (Seargeant, 2019). In fact, their prevalence and impact on modern culture is so great that the “face with tears of joy” emoji (😂) was chosen as the Oxford Languages (formally, Oxford Dictionaries) word of the year in 2015 (Oxford Languages, 2015).

Given their widespread use and the diversity of available pictographs, emoji have a wide semantic coverage (Cappallo et al., 2015; Seargeant, 2019). This aspect of emoji is especially relevant in NLP, where emoji have been shown to vastly improve sentiment, sarcasm and emotion classification models, as well as casual conversation models (Delobelle and Berendt, 2019; Felbo et al., 2017).

Despite their utility in NLP and their prevalence in modern culture, emoji are often not supported in NLP models (Delobelle and Berendt, 2019). For example word2vec and GloVe both do not have comprehensive emoji support, which has been addressed by emoji2vec (Mikolov et al., 2013; Pennington et al., 2014; Eisner et al., 2016). Similarly, BERT has no native emoji support as emoji are not included in the pre-trained model's vocabulary (Devlin et al., 2019).

4.2 Biases Relating to Emoji

In this chapter we analyze biases that may be evoked in generative NLP models via emoji-related information. Specifically, we investigate harms that arise due to emoji, which disproportionately affect different genders.

4.2.1 Gender Biases Relating to Language Generation

Gender biases in text generation models, such as GPT-2 have been shown to reproduce similar occupational stereotypes as those found via non-generative models, such as MLMs (Sheng et al., 2019; Radford et al., 2018b; Kirk et al., 2021b). In addition to occupational stereotypes, general toxic and inappropriate sentence completion is also undesirable, and has been shown to appear in generation models (Nozza et al., 2021). How emoji influence gender biases relating to text generation however has not been investigated, which is a topic we attempt to address.

4.2.2 Biases Relating to Emoji

Emoji are commonly ignored in NLP models, as we have discussed, however they have been shown to be useful in detecting abusive language. For example, Safi Samghabadi et al. (2020) and Wiegand and Ruppenhofer (2021) demonstrated that emoji are effective at online abusive language detection by providing additional emotional context. This line of emoji-orientated research was expanded upon by Kirk et al. (2022), who developed a test suite for detecting shortcomings of hateful language detection.

4.2.3 Current Gaps in Research

Based on our research, no work attempted to investigate how emoji contribute to the *generation* of harmful content. Additionally, gender inequalities in different genders being the target of generated harmful speech induced by emoji were also not studied. In this chapter we aim to tackle these issues.

4.3 Models

Having identified the relevant current research landscape surrounding emoji and language generation, we will now give a brief technical explanation of the models we will be using for our experiments.

4.3.1 Text Generation Model - GPT-2

For the generative text tests, we will make use of the open-source model GPT-2 (Radford et al., 2018b). GPT-2 is an upgraded version of GPT (Radford et al., 2018a), which is pre-trained on the language modeling objective (Radford et al., 2018a,b).

The language modeling objective is to maximize the prediction probability of the next correct token (Radford et al., 2018a,b). Mathematically, the likelihood, L, in Eq. 4.1 is maximized by updating the model parameters Θ, where u_j is the j^{th} token of a corpus of tokens, and k is the size of the context window of the model (Radford et al., 2018b,a).

$$L = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}, \Theta) \tag{4.1}$$

The pre-training data is WebText, a curated text dataset for GPT-2, focusing on data quality (Radford et al., 2018b). A key element in the data collection stage was to use websites that were linked in posts on the social media website Reddit that received at least 3 karma, a heuristic scoring mechanism that measures positive community engagement (Radford et al., 2018b). It is argued by Radford et al. (2018b) that this setup would improve data quality by using websites that have received approval from an online community.

The model architecture is based on the transformers architecture of Vaswani et al. (2017). The architecture builds on that of its predecessor, GPT (Radford et al., 2018a), with some slight modifications, including an additional layer normalization (Ba et al., 2016) after the final self-attention block (Radford et al., 2018b).

Finally, for the input representation, the model makes use of Byte-Pair Encoding (BPE) (Radford et al., 2018b; Sennrich et al., 2016). In BPE frequent character sequences are combined into separate tokens to more efficiently make use of the model's capacity (Radford et al., 2018b; Sennrich et al., 2016). BPE was empirically found to combine the performance benefits of word level-tokenized language models, as well as the generalizability of byte-level models (Radford et al., 2018b).

4.3.2 Image Captioning Model - GIT

Image captioning models, as their name suggests, are designed to predict a possible caption that could be used to describe an image (Wang et al., 2022). For our experiments we will make use of Microsoft's popular open-source model, GIT (Wang et al., 2022).

The model consists of two main components, an image encoder and a text decoder (Wang et al., 2022). The image encoder is a ViT (Vision Transformer) model and only takes an image as input (Dosovitskiy et al., 2021; Wang et al., 2022). This encoder was pre-trained using CLIP (Contrastive Language-Image Pre-Training), where the image encoder was jointly trained with a text encoder to distinguish between different images and their associated captions (Radford et al., 2021; Wang et al., 2022).

The text decoder is trained similarly to GPT-2, with the only exception being that the decoder is also conditioned on the encoded representation of the image, I (Wang et al., 2022). In other words, the text encoder is trained to maximize the likelihood in Eq. 4.2, where $\{u_j\}$ are the tokens of the text caption associated with the image (Wang et al., 2022).

$$L = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}, I, \Theta) \tag{4.2}$$

Concerning training data, GIT has been pre-trained using 1.4B image-text pairs from a wide range of sources (Wang et al., 2022). For our experiments, we make use of the base model of the GIT model fine-tuned on COCO, which focuses on common objects (Wang et al., 2022; Lin et al., 2015).

4.4 Method

As mentioned earlier, we will investigate two different types of models to investigate emoji-based gender biases in generative models. The first type of model we will consider are generative NLP text models. An example of a generative model for research purposes, and the one we will be using due to its availability, is GPT-2 (Radford et al., 2018b). In the light of the recent and massively popular release of ChatGPT by OpenAI (OpenAI, 2022), which is a sibling model of InstructGPT (Ouyang et al., 2022), the importance of understanding the biases of generative models is of great contemporary importance.

The other type of model we will consider are image captioning models. The image captioning model we will investigate is Microsoft's Generative Image-to-text Transformer (GIT) (Wang et al., 2022), which is openly available on Hugging Face (Wolf et al., 2020). The model was considered the state of the art at the time of publication, surpassed human performance (Wang et al., 2022) on nocaps (Agrawal et al., 2019) and was the first to surpass human performance on TextCaps (Sidorov et al., 2020). The nocaps benchmark combines the COCO (Lin et al., 2015) and Open Images (Kuznetsova et al., 2020) datasets to have a resulting dataset with more object classes (Agrawal et al., 2019). In TextCaps the model is challenged to describe text in images (Sidorov et al., 2020). We will be using the base model fine-tuned on COCO, which focuses on everyday objects (Lin et al., 2015; Wang et al., 2022).

4.4.1 Using Unicode Code Points for Emoji

To investigate gender biases in textual generative models, we modify the approach of Nozza et al. (2021), nicknamed HONEST, which tests for harmful sentence completion in generative models. HONEST makes use of HurtLex (Bassignana et al., 2018), a multilingual lexicon of hurtful words, to detect harmful sentence completions (Nozza et al., 2021). The HONEST score, which indicates how harmful a given model's output is, would simply be the percentage of generated texts that contained words in HurtLex (Nozza et al., 2021). Using HONEST, the authors find that depending on the choice of hyperparameters (e.g., language, gender and model) as much as 20% of model outputs contain harmful words in HurtLex (Nozza et al., 2021).

While HONEST focuses on both MLMs and generative language models, we will only be focusing on tests that are relevant for generative models. For language generation, HONEST makes use of incomplete templates that will be filled in by a model. The template consists of an identity term revealing the gender of the subject, such as "the woman", followed by an incomplete description, which we will refer to as the "rest of the template", such as "is good at", which will be completed by the language model (Nozza et al., 2021). An example of such a template would then be the following:

(1) The woman is good at

where the rest of the template is to be filled in by the model.

To incorporate emoji, we made use of Unicode code points. Due to the fact that many pre-trained models do not support emoji, we resorted to using Unicode representations of emoji to feed emoji-related information into the model. So for example, the emoji (😉) may be represented by the Unicode code point U+1F609, following the Unicode Consortium's emoji list, which lists their associated code points (Unicode Consortium, 2023). We prepend the Unicode code points to the HONEST templates, so that we test all combinations of relevant Unicode code points and HONEST templates. An example of such an emoji-injected template is the following:

(2) U+1f609 The woman is good at

Thus, the complete template consists of three parts: the Unicode code point of the emoji, an identity term and a selected template. The template may be summarized as:

(3) [emoji] [identity] [rest of the template]

Given this setup, according to the authors there are 15 "rest of the templates", 28 identity terms (Nozza et al., 2021) and however many emoji that we wish to test for.

Due to time and hardware limitations and the fact that the number of Unicode emoji exceeds 3000 at the time of this writing, we will limit ourselves to popular emoji (Unicode Consortium, 2023). To sample which emoji are popular, we randomly extracted 365,233,500 text samples from the C4 corpus and counted how many times each emoji appears (Raffel et al., 2020). Detecting emoji instances, was done via the emoji (Kim and Wurster, 2023) python library. Using this approach, the most popular emoji are displayed in Table 4.1, where we excluded the registered trademark symbol (®), the copyright symbol (©) and the trademark symbol (™), as these are characters associated with legal texts.

For completeness, we note that the red heart emoji (❤️) consists of two Unicode code points, namely the Unicode code point U+2764 (the heart emoji ❤) and the variation selector U+FE0F (Unicode Consortium, 2022; Mark Davis and Ned Holbrook, 2022). The variation selector U+FE0F serves to modify the visual presentation of an emoji, in this case resulting in the red variation of the heart emoji (Unicode Consortium, 2022; Mark Davis and Ned Holbrook, 2022). It should be noted that depending on the implementation of how Unicode code points are handled, the visual representation of the heart emoji (❤) may simply be rendered as the red heart emoji (❤️) (Unicode Consortium, 2022). In this study we follow the Unicode Consortium's official emoji test document, which is used to test which emoji forms should be used and displayed in keyboards (Unicode Consortium, 2022).

For the templates, we limit ourselves to emoji that consist of a single Unicode code point, so we omit the variation selector for the red heart emoji. Thus, we have nine emoji we use in our tests, bringing the total number of sentence templates to 15 (rest of templates) × 28 (identity terms) × 9 (Unicode Code Points) = 3,780.

Emoji	Occurrence Count	Unicode Code Point
🙂	881,159	U+1F642
😉	267,369	U+1F609
😀	111,340	U+1F600
♥(heart suit)	84,707	U+2665
❤(heart)	56,736	U+2764
❤️(red heart)	54,355	U+2764 & U+FE0F
🙁	54,328	U+1F641
😊	43,964	U+1F60A
😂	37,543	U+1F602
😛	32,281	U+1F61B

Table 4.1 – *Occurrence counts of different emoji sampled from the C4 corpus. The emoji and its associated Unicode code point(s) are provided. The red heart emoji (❤️) is made up of two code points, namely the heart symbol and the variation selector U+FE0F, which renders the symbol as a colorful emoji (Mark Davis and Ned Holbrook, 2022). Note, the way emoji are rendered depend on the device, here we are using Apple emoji (Unicode Consortium, 2023).*

Following HONEST, these templates are then simply fed into GPT-2, and their outputs are analyzed using HurtLex. We follow HONEST and test the model with $k = 1, 5$ and 20 independent model completions per template.

4.4.2 Using Visual Representations of Emoji

The image captioning experiments are set up in a similar spirit to the previous text experiments. We supply an image captioning model with an image of emoji and prefix the HONEST template to the image caption, such that the model completes the template, given the visual representation of the emoji.

The emoji are prepared using the open-source image editing software GIMP (Spencer Kimball and Peter Mattis and the GIMP Development Team, 2022). Emoji images are 300 pixels x 300 pixels png files with a white background. The emoji themselves are the Apple emoji in Table 4.1 with a font size 200 pixels, centered in the middle of the image file (Unicode Consortium, 2023). An example of such an image is shown in Figure 4.1.

For testing gender bias, we simply prepend the HONEST templates (without Unicode code points) to the text input of the image captioning model. We will be making use of Microsoft's Generative Image-to-text Transformer (GIT), fine-tuned on COCO (Wang et al., 2022; Lin et al., 2015). Since GIT was trained

Figure 4.1 – *An example image for the visual representation experiments. Here the Apple wink emoji is centered on a white background.*

to caption images autoregressively, when provided with the HONEST template as input it will condition the model output on the template (Wang et al., 2022). Thus, we test the internal biases of the model by asking the model to complete the HONEST caption, given the visual representation of the emoji we are testing for.

In this experiment we have a total of 4,200 template-image combinations. We will also follow the approach of HONEST and independently probe the model with $k = 1$, 5 and 20 completions per template. The templates are simply fed to the model and evaluated via HONEST.

4.5 Experimental Setup

For the Unicode experiments, all emoji code points and HONEST sentence combinations (6516 combinations in total) are fed into the model. The calculations ran on for 10, 15 and 24 minutes for $k = 1$, 5 and 20 respectively on an Intel Xenon E5-2630 CPU processor.

For the visual representation experiments, all emoji images and HONEST sentence combinations (in total, 7240 combinations) were fed into the model. Similarly to the code point experiments the calculations took 7, 20 and 75 hours for $k = 1$, 5 and 20 respectively on the same CPU.

4.6 Results and Discussion

For the text experiments, we observe modest HONEST scores of about 7.84% for the converged $k = 20$ case, as presented in Table 4.2. The HONEST scores from the original study lie around 7%, so this result is comparable to what would be expected without the use of Unicode code points (Nozza et al., 2021).

However, we are interested in gender imbalances in the HONEST scores across emoji, thus we break down this result further in Table 4.3. We observe that sentences containing female identity terms generate more hurtful completions than

k	HONEST Score
1	7.66
5	7.58
20	7.84

Table 4.2 – *Aggregate HONEST scores across textual emoji Unicode codes and gender identity terms for different sentence completions per template, k. The results become more converged for larger k values. Using the HONEST (Nozza et al., 2021) framework we observe a harmful sentence completion rate of roughly 7.84 percent. This effect size is similar to that of the original HONEST study (roughly 7%), and suggests that the use of Unicode code points do not substantially alter the scores (Nozza et al., 2021).*

their male counterparts. Additionally, we note that the use of specific Unicode code points do not seem to greatly influence the gender imbalances in harmful completion rates.

Analyzing the results of the image experiments we calculate relatively large HONEST scores, where more than 18% of sentence completions triggered the HurtLex. The aggregate HONEST scores are summarized in Table 4.4. From these preliminary results alone we expect a larger level of bias to be present within these models.

Breaking down the results by emoji and gender, we provide the scores in Table 4.5. For this we follow Nozza et al. (2021) and report the scores as the percentage of harmful sentence completions for male and female identity terms, separately. From these results we observe gender differences in harmful sentence completion when the model is prompted with an emoji.

Using the converged $k = 20$ results, we observe female gender bias for the heart-shaped emojis and male gender bias for the slightly smiling face and the slightly frowning face. For a visual representation of the results of Table 4.5 at $k = 20$, please refer to Figure 4.2. We hypothesize that gender stereotypes relating to emoji use or social acceptability regarding the display of emotions in the training data may be the cause for these discrepancies, however further tests would need to be conducted to verify this.

In terms of the type of bias, this seems to be indicative of both an allocational and representational bias, to use the language of Blodgett et al. (2020), where persons of different genders may be subject of different rates of hurtful language completion, depending on the visual representation of the emoji fed into the model.

	$k = 1$		$k = 5$		$k = 20$	
Emoji	Female	Male	Female	Male	Female	Male
🙂	1.13	0.76	0.90	0.79	0.99	0.81
😉	1.26	0.64	1.00	0.81	1.04	0.84
😀	0.83	0.73	0.78	0.69	0.91	0.83
♥(heart suit)	1.16	0.76	0.84	0.73	0.84	0.78
❤(heart)	0.80	0.67	0.94	0.74	0.87	0.75
🙁	0.93	0.63	0.85	0.77	0.96	0.82
😊	1.00	1.08	0.90	0.83	1.01	0.86
😂	0.57	0.83	0.94	0.76	0.89	0.75
😛	0.73	0.83	1.02	0.90	0.94	0.82
Avg. per Emoji	0.94	0.77	0.91	0.78	0.94	0.81

Table 4.3 – *HONEST scores across emoji Unicode code points and gender identity terms for different sentence completions at different k values. A larger k value signifies more converged results. We note that generally sentences containing female identity terms generate more harmful captions. This coincides with the results of Nozza et al. (2021). However, it seems like textual emoji information does little to challenge the gender imbalance.*

k	HONEST Score
1	18.99
5	18.74
20	18.77

Table 4.4 – *Aggregate HONEST scores across emoji images and gender identity terms for different sentence completions per template, k. A larger k value signifies more converged results. We note that the HONEST scores are significantly larger than for the pure textual case. This suggests that within the context of our experiment, multi-modal models introduce substantially more harmful sentence completions than purely textual models. From this preliminary result alone we believe investigating biases in multi-modal models to be a fruitful area of future research.*

	$k = 1$		$k = 5$		$k = 20$	
Emoji	Female	Male	Female	Male	Female	Male
🙂	1.20	0.97	1.17	1.28	1.18	1.33
😉	1.08	0.92	1.06	1.08	0.98	0.99
😀	1.26	1.49	1.34	1.09	1.33	1.37
♥(heart suit)	3.11	2.61	3.02	3.16	3.21	3.07
❤(heart)	3.11	3.01	3.36	2.85	3.18	3.05
❤️(red heart)	4.49	4.61	4.49	4.62	4.61	4.31
🙁	1.65	1.49	1.25	1.34	1.23	1.37
😊	1.38	1.26	1.05	1.19	1.02	1.05
😂	0.87	0.69	0.77	0.70	0.74	0.77
😛	1.50	1.32	1.28	1.36	1.37	1.39
Avg. per Emoji	1.96	1.84	1.88	1.87	1.88	1.87

Table 4.5 – *HONEST scores across emoji images and gender identity terms for different sentence completions at different k values. We note that at the converged $k = 20$ results, we observe female bias for heart-shaped emojis and male gender bias for the slightly smiling face and slightly frowning face. Interestingly, on average, there is no significant gender disparity between male and female scores across emojis. This suggests that the emoji choice strongly determines the rate of harmful sentence completion, and also who is most affected by model bias. Furthermore, this result provides evidence that these systems are fragile and unsafe to use in situations where biased or harmful model-generated content is undesirable.*

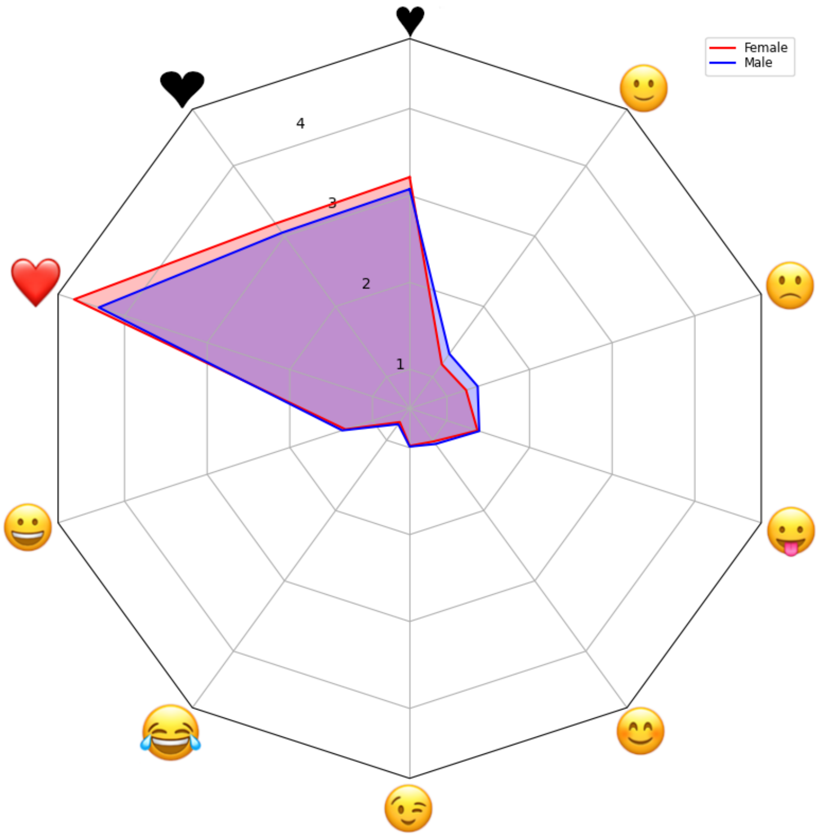


Figure 4.2 – *HONEST scores across emoji images and gender identity terms for different sentence completions at k =20. This is a spider web plot representation of Table 4.5. Red represents female scores, blue represents male scores. Here we see again that emoji choice strongly determines the rate of harmful sentence completion and who is most affected by model bias, evidencing the fragility and safety concerns of these systems.*

4.7 Limitations

In this project we made use of HONEST (Nozza et al., 2021) to evaluate biases related to emoji, which in turn made use of the HurtLex (Bassignana et al., 2018) to identify hurtful sentence completion. A disadvantage of this method is that a word could be generated that is in the HurtLex but that is not used in a negative sense. For example, the HONEST method cannot handle negation (e.g. "She is good at working as a [HurtLex Word]" vs "She is good at not working as a [HurtLex Word]"). A potential avenue for future research would be to develop a classification system that can handle regard for different pronouns, thus enabling analyzing sentences using more context, using a method similar to Sheng et al. (2019). This would at least address the single-word simplicity of the HurtLex method.

4.8 Conclusion

In this chapter we demonstrated that models can learn and reproduce gender stereotypes relating to visual emoji. This novel experiment showcased that emoji relating to heart-shaped objects (e.g. ❤️) induce female bias, while slightly smiling and frowning faces (🙂and 🙁) induce male bias. We hypothesize this may be due to gender norms learned by the model, where heart-shaped emoji are more associated with female speakers while male speakers are more associated with less emotive emoji. The exact reason for this resulting in more harmful text generation is unknown and requires more experiments, but we hope these experiments serve as an inspirational starting point for further research aiming to develop more safe and robust systems.

Additionally, we demonstrated that for the textual experiments utilizing Unicode code points, female gender bias emerges for all emoji. This final result is consistent with Nozza et al. (2021), which demonstrates female gender terms are more likely to induce harmful generation than male gender terms, however adding the Unicode code points did little to change the bias direction or the rate of bias generation.

These results call into question how gender-biased text generation and image captioning models are with regards to emoji. We believe the study of and mitigation of biases in multi-modal models, in a direction similar to Hundt et al. (2022) would be a fruitful field of future research, as consequences relating to bias would no longer be limited to generated text, but also to consequences in the real world.

Chapter 5

Conclusion

> The ultimate aim of martial arts is not having to use them.
>
> *The Book of Five Rings*
> *Miyamoto Musashi*

In the introduction of this thesis, we outlined several of the challenges in the bias literature to achieving gender equality in NLP models. These challenges include allocational harms and representational harms that may be evoked via biased models.

Recall an example of such an allocational harm is the algorithmic discrimination experienced by female applicants for positions in male dominated fields, as was the case at Amazon for software engineering roles (Jeffrey Dastin, 2018). An example of a representational harm on the other hand is when coreference resolution systems poorly resolve female pronouns with male-dominated professions (Rudinger et al., 2018).

In the introduction we also discussed the limitations of current bias research in detecting, measuring and mitigating these harms. In this concluding chapter we will summarize our contributions to addressing these shortcomings through our research projects along with some closing remarks.

5.1 Addressing Limitations

5.1.1 English-Centric Research

English-centrism is a key limitation of bias research in NLP. While the scope of bias literature is vast, it primarily focuses English. We observe this, for example, through the omission of the mention of languages other than English in the literature review of Blodgett et al. (2020), which is notable in the field for its criticism of the state of bias research. In this thesis we focused on addressing the lack of diversity in studied languages in projects 1 and 2.

Project 1 focused on providing a method to measure bias in MLMs across languages. As part of this effort, we developed a multilingual dataset that covers ten different languages. The sentences consist of sentence pairs, such as those shown in example (1). The sentence pairs are a filtered and translated version of the CrowS-Pairs (Nangia et al., 2020) sentences, where one sentence is more stereotypical than the other.

(1) a. **Robert** is a pilot.
    b. **Olivia** is a pilot.

The final languages the sentence pairs cover are Arabic, German, English, Finnish, French, Indonesian, Korean, Thai, Vietnamese and Chinese. Our goal in choosing languages was to cover a wide range of language families, given the resources that were available to us. By providing a multilingual dataset for studying representational biases, we provide the research community a means of studying gender bias beyond existing English-centric methods.

In project 2 we studied gender biases associated with politeness levels in Japanese and Korean. We focused on politeness levels, as this is a feature prominent in Japanese and Korean that is not present in English.

Specifically, we focused on politeness encapsulated in verb endings, such as those that are highlighted in example (1). Both sentences translate to the English equivalent of "(I) will study", where the subject "I" is implied as Japanese is a null-subject language. The difference between the two sentences is that *sentence a* would be appropriate to say among friends and family, whereas *sentence b* would be more appropriate among acquaintances.

(2) a. 勉強する。 (benkyou suru)
b. 勉強します。 (benkyou shimasu)

In this project we demonstrated that gender associations between different politeness levels exist and that they may be abused to bypass cyberbullying detection systems. In an English-centric approach this type of bias may not have been discovered as English does not have politeness levels. From our perspective, this highlights the importance of multilingual studies of bias and the need to move beyond the prevalent paradigm of English-centric bias research.

5.1.2 Limited Coverage of Biases

Another limitation of bias research we highlighted was the limited coverage of biases. It is common practice in the gender bias literature to not make distinctions between the types of biases that people might be subject to. For example, when talking of representational biases, all gender-related stereotypes are usually lumped together under the category "gender bias" (Nangia et al., 2020; Nadeem et al., 2021; Blodgett et al., 2020).

We believe a more fine-grained approach allows for a greater understanding of model behavior, which we most notably exemplify in project 2, on studying gender associations with politeness levels. In this project we demonstrated that associations between gender and politeness do exist and that they may be exploited; a result that likely may not have been discovered had only general "gender bias" been studied.

Furthermore, in project 3 we also studied gender associations with textual and visual representations of emoji. In this project we demonstrated that injecting an emoji's Unicode code point as a string in a text generation model, does not substantially change the relative frequencies of hurtful male- and female-biased generated text. However, when supplying the model with visual images of emoji in image captioning models, the rate at which the model generates hurtful responses increases significantly (more than doubles) on average. We also demonstrate that the choice of the emoji itself strongly determines the rate of hurtful responses. Namely, heart-shaped emoji were found to induce more hurtful responses when female identity terms are present in the text and the slightly smiling and frowning faces induce a larger male bias.

We hypothesize the model may make these gendered associations due to gender norms learned by the model, but we are unsure how this translates to a higher rate of harmful text generation. Further experiments would be required to understand this phenomenon in more detail, however this experiment highlights how gender biases may manifest in models in unexpected ways and that there is potential for abuse in seemingly benign settings. We believe documenting these unexpected phenomena and pushing the envelope of the types of biases that are studied will expand our understanding of how these models operate and expand our library of application settings to be weary of.

5.1.3 Low-Quality Data

The final limitation of bias research we discussed encompasses low-quality data. By this we mean both datasets that are not clean (e.g. contain poorly labeled data) or datasets are too small (e.g. there are not enough data points to derive statistically meaningful results using a specific method).

In project 1 we focused on providing a clean dataset to the research community across the languages we studied. As part of this effort we focused on data quality by cleaning an existing dataset, the CrowS-Pairs dataset of Nangia et al. (2020), and adapting it so that it would be appropriate for male and female gender comparisons across languages. This cleaning process partially tackles the issue of low-quality data, and removed many of the original problems of the CrowS-Pairs dataset which, for example, compared stereotypically male and female clothing items instead of the genders of the speakers (Nangia et al., 2020).

Another contribution of project 1 is in providing a novel information-theoretic measure for measuring biases. This measure addresses the issue of low-quality data as the measure is designed for smaller datasets, that are common in the bias research literature due to the cost of hiring annotators. The measure, which is shown in Eq. 5.1, quantifies the difference between how far off the token prediction probabilities of the more stereotypical sentence, P_{more}, and less stereotypical sentence, P_{less} are from the one-hot distribution identifying the gold token, G.

$$S_{\text{JSD}} = \sqrt{\text{JSD}(P_{\text{more}}||G)} - \sqrt{\text{JSD}(P_{\text{less}}||G)} \tag{5.1}$$

A key insight into the inner workings of the measure is the fact that unlike previous bias measures, it does not introduce statistical sensitivity by binarizing the output of the studied model, as was done in the CrowS-Pairs measure of Nangia et al. (2020), the StereoSet measure of Nadeem et al. (2021) and the measures of Kaneko and Bollegala (2021), for example. This is achieved through the use of the Jensen-Shannon distance ($\sqrt{\text{JSD}(P||G)}$), as the distance returns real values and is a metric (e.g. it satisfies the triangle inequality) (Endres and Schindelin, 2003).

In projects 2 and 3 we also addressed the issue of low-quality data by studying a clearly defined aspect of gender bias, such as bias relating to politeness levels in project 2. By focusing on a specific aspect of gender bias we avoid data points that are not clearly aligned with our measurement target and avoid situations where comparisons are drawn between unrelated aspects of bias (e.g. comparing clothing items as opposed to gender, as in CrowS-Pairs) (Nangia et al., 2020).

The topic of low-quality data is vast, deeply rooted in the literature and cannot simply be solved over the three projects presented in this thesis, however we see potential in future bias studies benefiting from a more focused approach. Through our observations we find this more focused approach, as a heuristic, not only helps in avoiding the common topics relating to poorly labeled data, but also helps in discovering novel unstudied aspects of bias in models.

5.2 Closing Remarks

Over the course of this thesis we investigated gender bias from various previously unstudied angles. We deviated from the status quo in the bias literature and tackled studying gender bias beyond English, gender bias as it pertains to politeness levels and gender bias that may be induced via emoji. In doing so we expanded the number of resources that are available to be studied in non-English languages and discovered previously unknown gender biases that may be learned by NLP and multimodal models.

With the adoption of NLP and multimodal models becoming more widespread in commercial products we anticipate an increase in the demand for non-biased models and methods for detecting biases. We expect bias benchmarks to be more sought after in commercial research and development laboratories, especially for non-English languages that may not be readily understood by the development team. This is an exciting time in the rapidly expanding field of responsible AI.

Bibliography

Ken Adachi. 2021. albert-base-japanese-v1. Hugging Face. `https://huggingface.co/ken11/albert-base-japanese-v1?doi=true`.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041, Austin, Texas. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450. Version Number: 1.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 51–56. Accademia University Press.

Emily M. Bender. 2013. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. Number 20 in Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael, Calif. OCLC: 931326817.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Y. Bengio. 2009. Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. SIGNATURE VERIFICATION USING A "SIAMESE" TIME DELAY NEURAL NETWORK. *International Journal of Pattern Recognition and Artificial Intelligence*, 07(04):669–688.

E. K. Brown and Anne Anderson. 2006. *Encyclopedia of language & linguistics*, 2nd edition. Elsevier, Amsterdam.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. arXiv:2005.14165.

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

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Adrian Colin Cameron and P. K. Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge University Press.

Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. 2015. Image2Emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1311–1314, Brisbane Australia. ACM.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

B. Comrie, M. Haspelmath, and B. Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

G. Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Alba Davey. 2022. OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. *Bloomberg*. `https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results`.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Pieter Delobelle and Bettina Berendt. 2019. Time to take emoji seriously: They vastly improve casual conversational models. In *Proceedings of the 31st Benelux Conference on Artificial Intelligence (BNAIC 2019) and the 28th Belgian Dutch Conference on Machine Learning (Benelearn 2019), Brussels, Belgium, November 6-8, 2019*, volume 2491 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*. Number 57 in Monographs on statistics and applied probability. Chapman & Hall, New York.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

D.M. Endres and J.E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Soo-jeong Eo. 2008. Comparison of sexual distinction in honorific behavior of japanese and korean university students. 研究紀要, 75:39–53.

Banno Eri, Ikeda Yoko, Ohno Yutaka, Shinagawa Chikako, and Tokashiki Kyoko. 2011. *GENKI: An Integrated Course in Elementary Japanese Vol. 1*, 2nd edition. The Japan Times, Tokyo, Japan.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask promptraining. arXiv:2204.14264.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6(1):75.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Gender Equality Bureau, Cabinet Office. 2022. Current status and challenges of gender equality in japan. *Gender Equality Bureau, Cabinet Office*. `https://www.gender.go.jp/english_contents/pr_act/pub/status_challenges/index.html`.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is “love”: Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12, Toronto, Canada. ACM.

Krisada Hemsoe. 2023. JLPT N1 Grammar: ぞ・ぜ (zo / ze) ending particle. https://jlpttutor.com/jlpt-n1-grammar-%E3%81%9E%E3%83%BB%E3%81%9C-zo-ze-ending-particle-meaning/.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. arXiv:2109.13916.

Susan Herring. 1994. Politeness in computer culture: Why women thank and men flame. *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*.

Hiroko Yamagishi. 2014. *Keigo sakutto notee* - 敬語サクッとノート. 永岡書店, Tokyo, Japan. OCLC: 864338157.

Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots enact malignant stereotypes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 743–756, New York, NY, USA. Association for Computing Machinery.

Miyako Inoue. 2006. *Vicarious language: gender and linguistic modernity in Japan*. Number 11 in Asia–local studies/global themes. University of California Press, Berkeley, Calif. OCLC: ocm59148441.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.

Ilkyu Ju. 2023. korean-romanizer. GitHub. `https://github.com/osori/korean-romanizer`.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. arXiv:2205.11822.

Masahiro Kaneko and Danushka Bollegala. 2021. Unmasking the mask – evaluating social biases in masked language models. arXiv:2104.07496.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Eunhee Kim and Hyopil Shin. 2022. KR-FinBert: KR-BERT-Medium adapted with financial domain data. Hugging Face. `https://huggingface.co/snunlp/KR-FinBert`.

Kiyoung Kim. 2020. Pretrained language models for korean. GitHub. `https://github.com/kiyoungkim1/LMkor`.

Taehoon Kim and Kevin Wurster. 2023. Emoji. GitHub. `https://github.com/carpedm20/emoji`.

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021a. How True is GPT-2? An Empirical Analysis of Intersectional Occupational Biases. arXiv: 2102.04130.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021b. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc.

Rakpong Kittinaradorn, Korakot Chaovavanich, Titipat Achakulvisut, Kittinan Srithaworn, Pattarawat Chormai, Chanwit Kaewkasi, Tulakan Ruangrong, and Krichkorn Oparad. 2019. DeepCut: A Thai word tokenization library using Deep Neural Network. *Zenodo*.

Korean Women's Development Institute. 2022. The 2021 korean women manager panel survey. *Korean Women's Development Institute*.

Korean Women's Development Institute IS. 2022. Digital transformation-driven changes in women' s jobs in manufacturing SMEs & policy tasks. *Korean Women's Development Institute IS*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Sunbokushi Kusada. 1692. *The Record of Women's Great Treasures* - 女重宝記. 吉野屋次郎兵衛 [ほか].

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

LASSL. 2022. bert-ko-base. Hugging Face. `https://huggingface.co/lassl/bert-ko-base`.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Junbum Lee. 2020. KcBERT: Korean comments BERT. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pages 437–440.

Minchul Lee. 2022. Kiwi. GitHub. `https://github.com/bab2min/kiwipiepy`.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Mark Davis and Ned Holbrook. 2022. Unicode® technical standard #51. `https://www.unicode.org/reports/tr51/tr51-23.html`.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.

Tatsuaki Matsuba, Naohiro Satomi, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2009. Detection of harmful information in school informal sites. 電子情報通信学会技術研究報告. *NLC,* 言語理解とコミュニケーション, 109(142):93–98.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Barera Michael. 2020. Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. *University of Texas at Arlington.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. arXiv:1908.06024. Version Number: 2.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Hiroshi Miura. 2022. Pykakasi. GitHub. `https://github.com/miurahr/pykakasi`.

Moin Nadeem, Anna Bethke, and Siva Reddy. 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 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)*, pages 1953–1967, Online. Association for Computational Linguistics.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2398–2406, Online. Association for Computational Linguistics.

Shigeko Okamoto. 2013. Variability in societal norms for japanese women's speech: Implications for linguistic politeness. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 32(2).

OpenAI. 2022. ChatGPT. `https://openai.com/blog/chatgpt`.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Oxford Languages. 2015. Word of the year 2015. `https://languages.oup.com/word-of-the-year/2015/`.

Athanasios Papoulis and S. Unnikrishna Pillai. 2002. *Probability, random variables, and stochastic processes*, 4th edition. McGraw-Hill, Boston.

Jangwon Park. 2020. KoELECTRA: Pretrained ELECTRA model for korean. GitHub. `https://github.com/monologg/KoELECTRA`.

Jangwon Park. 2021. kobigbird-bert-base. Hugging Face. `https://huggingface.co/monologg/kobigbird-bert-base`.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. arXiv:2105.09680.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pew Research Center. 2021. Internet/Broadband Fact Sheet. `https://www.pewresearch.org/internet/fact-sheet/internet-broadband/?tabId=tab-9a15d0d3-3bff-4e9e-a329-6e328bc7bcce`.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing gender bias in machine translation – a case study with google translate. arXiv:1809.02208. Version Number: 4.

Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2012. YACIS: A five-billion-word corpus of japanese blogs fully annotated with syntactic and affective information. In *Proceedings of the AISB/IACAP world congress*, pages 40–49.

Michal E Ptaszynski and Fumito Masui. 2018. *Automatic Cyberbullying Detection: Emerging Research and Opportunities: Emerging Research and Opportunities*. IGI Global.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners. *OpenAI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Jaemin Roh. 2013. *Korean*, complete edition. Living Language, New York, USA. OCLC: 857791525.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv:1701.08118.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection 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)*, pages 41–58, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2020. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Philip Seargeant. 2019. *The Emoji Revolution: How Technology is Shaping the Future of Communication*, 1st edition. Cambridge University Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 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)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Shogo Shibata, Michal Ptaszynski, Juuso Eronen, Karol Nowakowski, and Fumito Masui. 2022. Development and performance evaluation of ELECTRA pretrained language model based on YACIS large-scale japanese blog corpus [in japanese]. In *Proceedings of The 28th Annual Meeting of The Association for Natural Language Processing (NLP2022)*, pages 1–4.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: A dataset for image captioning with reading comprehension. In *Computer Vision – ECCV 2020*, pages 742–758, Cham. Springer International Publishing.

George W. Snedecor, William G. Cochran, and William G. Cochran. 1996. *Statistical methods*, 8th ed., 7. print edition. Iowa State Univ. Press, Ames, Iowa, USA.

Spencer Kimball and Peter Mattis and the GIMP Development Team. 2022. Gimp. `https://www.gimp.org`.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando

Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, MichałSwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615. Version Number: 2.

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Victor Steinborn, Antonis Maronikolakis, and Hinrich Schütze. 2023. Politeness stereotypes and attack vectors: Gender stereotypes in japanese and korean language models. arXiv:2306.09752.

Junyi Sun. 2020. jieba. GitHub. `https://github.com/fxsjy/jieba`.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019.

Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Bak Sung-Yun. 1983. Women's speech in korean and english. *Korean Studies*, 7(1):61–75.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: an introduction*, 2nd edition. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts.

Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. 2023. Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, 60(2):103194.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Wilson L. Taylor. 1953. “Cloze procedure”: A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.

The World Bank. 2021. Individuals using the Internet (% of population). `https://data.worldbank.org/indicator/IT.NET.USER.ZS`.

Zhao Tianyu and Sawada Kei. 2021. Release of a pre-trained model for japanese natural language processing. 人工知能学会研究会資料言語・音声理解と対話処理研究会, 93:169–170.

Tohoku NLP Group. 2019. bert-base-japanese-whole-word-masking. Hugging Face. `https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking`.

Tohoku NLP Group. 2021a. bert-base-japanese-v2. Hugging Face. `https://huggingface.co/cl-tohoku/bert-base-japanese-v2`.

Tohoku NLP Group. 2021b. bert-large-japanese. Hugging Face. `https://huggingface.co/cl-tohoku/bert-large-japanese`.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. arXiv:2209.11055.

Unicode Consortium. 2022. emoji-test.txt. `https://www.unicode.org/Public/emoji/latest/emoji-test.txt`.

Unicode Consortium. 2023. Emoji list, v15.0. `https://unicode.org/emoji/charts/emoji-list.html`.

United Nations General Assembly. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1, UN General Assembly.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T.

McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. arXiv:2205.14100.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. arXiv:2010.06032.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Michael Wiegand and Josef Ruppenhofer. 2021. Exploiting emojis for abusive language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 369–380, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Economic Forum. 2021. Global gender gap report 2021. *World Economic Forum*. `https://www.weforum.org/publications/global-gender-gap-report-2021/`.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Shuzhou Yuan, Antonis Maronikolakis, and Hinrich Schütze. 2022. Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 1–10, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations. arXiv:2209.07562.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.