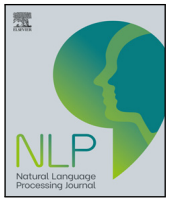
Natural Language Processing Journal 6 (2024) 100047

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



Gender bias in transformers: A comprehensive review of detection and mitigation strategies

Praneeth Nemani [a,d,*], Yericherla Deepak Joel [b], Palla Vijay [b], Farhana Ferdouzi Liza [c,1]

[a] Department of Computer Science and Engineering, IIIT Naya Raipur, Chhattisgarh, India
[b] Department of Data Science and Artificial Intelligence, IIIT Naya Raipur, Chhattisgarh, India
[c] School of Computing Sciences, University of East Anglia (UEA), Norwich, United Kingdom
[d] College of Engineering and Applied Science, University of Colorado Boulder, CO, USA

ARTICLE INFO

Keywords:
Gender bias
Transformer models
Bias mitigation
Binary gender assumption
Self-attention

ABSTRACT

Gender bias in artificial intelligence (AI) has emerged as a pressing concern with profound implications for individuals' lives. This paper presents a comprehensive survey that explores gender bias in Transformer models from a linguistic perspective. While the existence of gender bias in language models has been acknowledged in previous studies, there remains a lack of consensus on how to measure and evaluate this bias effectively. Our survey critically examines the existing literature on gender bias in Transformers, shedding light on the diverse methodologies and metrics employed to assess bias. Several limitations in current approaches to measuring gender bias in Transformers are identified, encompassing the utilization of incomplete or flawed metrics, inadequate dataset sizes, and a dearth of standardization in evaluation methods. Furthermore, our survey delves into the potential ramifications of gender bias in Transformers for downstream applications, including dialogue systems and machine translation. We underscore the importance of fostering equity and fairness in these systems by emphasizing the need for heightened awareness and accountability in developing and deploying language technologies. This paper serves as a comprehensive overview of gender bias in Transformer models, providing novel insights and offering valuable directions for future research in this critical domain.

1. Introduction

Artificial intelligence (AI) is often perceived as a neutral entity. However, as AI is created by humans, it reflects our prejudices, including gender bias (Nadeem et al., 2020, 2022). There is a growing concern in both the scientific community and the general public about the demographic biases in some AI applications (Schwartz et al., 2022). Gender bias can perpetuate harmful stereotypes and biases, leading to unfair treatment and discrimination. This bias can also limit opportunities for marginalized groups, particularly in areas such as employment, education, and healthcare. Furthermore, it can negatively impact the accuracy and fairness of natural language processing (NLP) applications, affecting the user experience and reliability of these systems. As NLP and machine learning (ML) tools gain prominence, it is becoming increasingly important to understand how they contribute to the formation of societal prejudices and preconceptions. NLP models are effective at modeling many different applications, but they can reinforce gender prejudice that is present in text corpora. Despite the fact that gender bias in NLP has been studied since the 90s (Baldwin et al., 1995; Ozieblowska, 1994), there is still much to learn about certain aspects. Therefore, it is crucial to examine how current research in NLP is establishing an entirely new field (Wang and Redmiles, 2019). To determine whether these studies are heading in the right direction to solve the problem, it is essential to evaluate whether they present scalable evaluation techniques and whether their objectives are well-stated. Bias refers to an unfair opinion or preference held in favor of or against a specific person or group (Simundic, 2013). In machine learning, bias can be caused by faulty assumptions in the algorithm or systemic prediction errors caused by the characteristics of the training data (Turney, 1995; Mehrabi et al., 2021). Gender bias, which favors or stereotypes one gender over another, particularly males over females, has been studied extensively in relation to NLP applications. Different researchers have proposed various definitions and inferences of gender bias and its impact on NLP.

1.1. Gender bias in word embeddings

Word embeddings are crucial in transformer-based NLP models. They are a fundamental component that helps transformers understand

* Corresponding author at: Department of Computer Science and Engineering, IIIT Naya Raipur, Chhattisgarh, India.
E-mail addresses: praneeth19100@iiitnr.edu.in, Praneeth.Nemani@colorado.edu (P. Nemani), yericherla19102@iiitnr.edu.in (Y.D. Joel), palla19102@iiitnr.edu.in (P. Vijay), F.Liza@uea.ac.uk (F.F. Liza).

[1] Fellow, IEEE.

https://doi.org/10.1016/j.nlp.2023.100047
Received 5 August 2023; Received in revised form 6 October 2023; Accepted 9 December 2023

and represent the meaning of words in a way that machines can process. Word embeddings are dense vector representations of words that capture semantic and syntactic relationships between them. These vectors encode contextual information about words based on their co-occurrence patterns in large text corpora. In transformer models, such as the popular architecture known as BERT (Bidirectional Encoder Representations from Transformers), word embeddings are used as input representations for the model. These embeddings provide the initial understanding of individual words within the context of a given sentence or document. The transformer model then processes these word embeddings through self-attention mechanisms, enabling it to capture contextual relationships between words. The attention mechanism allows the model to weigh the importance of each word in the context of the entire sentence, which is crucial for understanding the meaning of a word based on its surrounding words. The ability of word embeddings to capture semantic relationships allows transformer models to perform a wide range of NLP tasks effectively. These tasks include text classification, named entity recognition, sentiment analysis, machine translation, question answering, and many others. By leveraging word embeddings, transformers can learn and generalize patterns from vast amounts of text data, improving performance in various NLP tasks. The Artificial Intelligence and Emerging Technology Initiative of The Brookings Institution has explored the issue of bias and NLP research in their **"AI and Bias"** series, with a particular focus on gender bias. The author of one 2021 essay draws on their previous research at Princeton University's Center for Information Technology Policy, where they found that machine learning algorithms processing word embeddings can pick up biases similar to those of humans from the word associations in their training data (Brunet et al., 2019; Papakyriakopoulos et al., 2020; Zhao et al., 2019). One example of gender bias in word embeddings is the association of certain professions with gender. For instance, the word "nurse" may be more closely associated with the female gender than the male gender in a word embedding model. Similarly, "engineer" may be more closely associated with the male gender than the female gender. This bias can manifest in natural language processing applications such as automated resume screening.

Another example of gender bias is the case of **Amazon** (Anon, 2023). In 2018, it was reported that Amazon had developed an AI-powered hiring tool that was trained on resumes submitted to the company over a 10-year period. The tool was designed to screen resumes and rank candidates based on their qualifications, with the goal of identifying top candidates more efficiently. However, the tool was found to have a significant gender bias. This was because the tool was trained on resumes submitted to Amazon over the past decade, which were predominantly from men due to the tech industry's gender imbalance. As a result, the tool learned to associate certain words and phrases with male candidates and would downgrade resumes that contained language associated with women, such as references to women's colleges or women's sports teams. Amazon ultimately abandoned the AI-powered hiring tool, recognizing that it was not effective and potentially harmful due to its gender bias. The case highlighted the importance of ensuring that automated tools used in hiring and other applications are free from bias and trained on diverse datasets to ensure fairness and accuracy. To describe this representation professionally, one could say that Fig. 1 presents a conceptual illustration of the way in which gender bias can manifest in word embeddings.

1.2. Gender bias in machine translation

Gender bias in machine translation refers to the phenomenon where machine translation systems produce translations that reflect or reinforce gender stereotypes or where translations are inaccurate or inappropriate because of gender-related differences in the source and target languages (Stanovsky et al., 2019; Prates et al., 2020). One common example of gender bias in machine translation is the use of gendered pronouns. Many languages, such as French and Spanish,

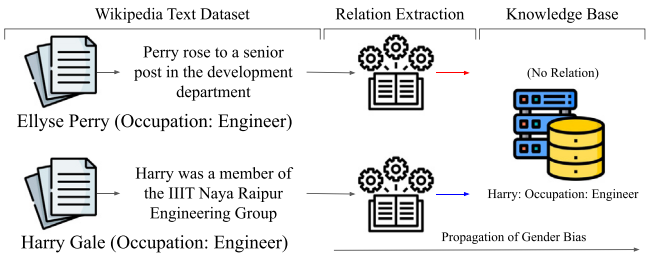


Fig. 1. Gender Bias in Word Embeddings.

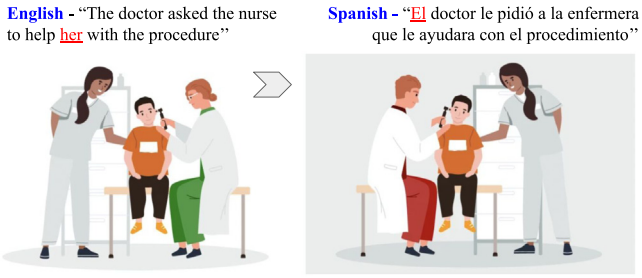


Fig. 2. Evidence of Gender Bias in MT even due to the presence of unambiguous gender context.

use gendered pronouns to refer to people, and machine translation systems may struggle to accurately translate sentences that contain these pronouns. For example, the French sentence *Le médecin a vu la patiente* can be translated into English as *The doctor saw the patient* or *The doctor saw the female patient*, but machine translation systems may default to using the masculine pronoun "he" instead of "she" when translating the sentence into English.

Another example of gender bias in machine translation is the use of gender stereotypes in translations. For instance, a machine translation system may translate a sentence like *She is a doctor* into a language where the word for doctor is masculine by default, resulting in a translation that reinforces the stereotype that doctors are male. Similarly, a system may translate a sentence like "He is a nurse" into a language where the word for nurse is feminine by default, which could be seen as reinforcing the stereotype that nurses are female. MIT press has suggested that the weight of prejudices and stereotypes, as well as the presence of equal gender weightage, can be used to assess gender bias in MT (Savoldi et al., 2021). In cases where a language with limited or no gender (such as English) is being interpreted into a language with significant grammatical gender, an ideal MT model should accurately translate and express the genders of words lacking gender in the input language (such as Spanish). Fig. 2 visually illustrates such a bias in the translation from English to Spanish.

1.3. Gender bias in caption generation

Caption generation generates a textual description that accurately describes the content of an image or a video. Gender bias can also manifest in caption generation, where the generated captions can be biased towards a particular gender. The issue of gender bias is not limited to machine translation and word embeddings; it can also be found in caption generation tasks. Tang et al. (2020) conducted a study and found that captioning datasets, such as the COCO dataset, may lead to unintentional gender-biased models due to intrinsic memorization. The COCO training dataset exhibits a significant gender bias, which makes the 3:1 male-to-female ratio in the dataset even more unbalanced. An ideal model should not identify a person as a woman based on the background of a house, for instance. Instead, an unbiased model should predict gender terms based on visual characteristics associated with the

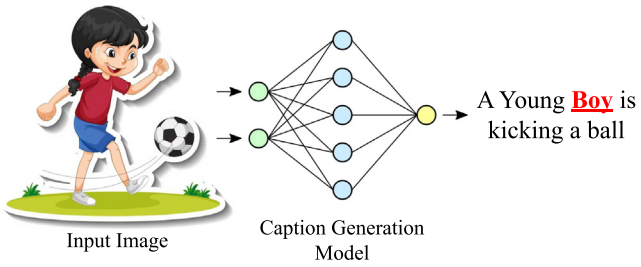


Fig. 3. Biased scenario of caption generation.

depicted individual. The problem of gender bias in caption generation is depicted in Fig. 3, where the model predicts the output based on biased data. Other studies, such as Hendricks et al. (2018), Hirota et al. (2022), also highlight the issue of bias in caption generation.

1.4. Gender bias in sentiment analysis

Sentiment analysis, which involves the automated identification of emotions or attitudes in large datasets, is a popular application of NLP and ML techniques. However, as pointed out by Thelwall (2018), this task can be susceptible to bias due to under-representation. In this study, the researchers investigated whether biases exist in the accuracy of lexical sentiment analysis when applied to reviews written by individuals of different genders. Specifically, the analysis focused on TripAdvisor reviews of hotels and restaurants in the UK, authored by UK residents. The aim was to compare the effectiveness of lexical sentiment analysis in detecting sentiments expressed by males and females. The study's results revealed that detecting sentiment in reviews written by males was more challenging than those written by females. This difficulty stemmed from the fact that male sentiment tended to be less explicit or overtly expressed. Furthermore, the researchers found no evidence to support the notion that gender-specific lexical sentiment analysis could effectively address this issue. Similarly, studies by Asyrofi et al. (2022) and Kiritchenko and Mohammad (2018a) have also highlighted the issue of gender bias in sentiment analysis. In their work, Asyrofi et al. (2022) introduced BisaFinder, an approach designed to uncover biased predictions in sentiment analysis systems using metamorphic testing. BisaFinder incorporates several key components, including the automatic generation of appropriate templates based on text fragments sourced from a large corpus. Kiritchenko and Mohammad (2018a) introduced the **Equity Evaluation Corpus (EEC)**, which comprises 8,640 English sentences meticulously selected to uncover biases pertaining to specific races and genders. The primary objective of their research is to investigate 219 automatic sentiment analysis systems that participated in the SemEval-2018 Task 1 'Affect in Tweets' shared task, utilizing the EEC dataset. The researchers' analysis reveals that several sentiment analysis systems exhibit statistically significant biases. Specifically, these systems consistently generate slightly higher sentiment intensity predictions associated with one race or gender.

The pervasive nature of gender bias in AI and its downstream ramifications is increasingly becoming evident through real-world instances. For instance, voice assistants often defaulted to female voices, and have not only reinforced subservient stereotypes but, in their earlier versions, also responded inappropriately to derogatory comments West et al. (2019)). A more critical manifestation of this bias was observed in health applications. Certain health apps demonstrated a bias towards recognizing symptoms described predominantly in male-associated language, potentially jeopardizing timely and accurate medical interventions for women. Lastly, the financial sector was not immune either; there were reports of automated credit systems favoring men over women despite the latter's superior financial records (Sweeney, 2013). These examples emphasize the consequential impact of unmitigated gender bias in AI systems across diverse sectors.

1.5. Representation of gender bias

The issue of gender bias in various NLP tasks is a matter of significant concern. In this context, Caliskan et al. (2017) demonstrated that biases could range from morally neutral attitudes, such as those towards flowers and insects, to more contentious ones related to race and gender. Furthermore, they could even provide factual representations, such as the distribution of gender across different professions or the commonness of first names. What is remarkable is that these biases are integrated into the semantics grasped by machine learning, emphasizing the influence of the data these models are trained on. Researchers have identified three primary categories of bias: **Denigration (A1)**, **Stereotyping (A2)**, and **Under-representation (A3)**.

Denigration involves the usage of racial, ethnic, or religious slurs, which can often be observed as a prevalent method of cyberbullying. This type of bias manifests as derogatory language aimed at demeaning specific groups as outlined by Waseem and Hovy (2016), who studied the prevalence of hate speech on Twitter. **Stereotyping** refers to individuals' cognitive representation of a particular social group. Tan and Celis (2019) showcased that word embeddings can perpetuate gender stereotypes, associating female terms more with family roles and male terms with career roles. From a geometrical perspective, the research revealed two primary insights. Firstly, the gender bias within these embeddings could be identified along a specific directional axis. Secondly, the embeddings revealed that words without a direct gender association could still be distinctively separated from words that defined gender. **Under-representation** pertains to the absence of identifiable group members from representative bodies and well-being indicators in proportion to their population's size. Researchers like Buolamwini and Gebru (2018) highlighted this form of bias in facial recognition software, showing differences in gender classification accuracy across skin tones. The researchers have introduced a methodology to assess biases in automated facial analysis algorithms and datasets, particularly concerning phenotypic subgroups. Leveraging the widely accepted Fitzpatrick Skin Type classification system, which is endorsed by dermatologists, the team evaluated the gender and skin type distribution in two major facial analysis benchmarks, and their analysis revealed a significant skew in these datasets towards lighter-skinned individuals. Researchers have extensively studied these categories of bias, as mentioned by Zhao et al. (2018), where the authors discussed gender biases in machine-translated content. By recognizing and categorizing different types of bias, researchers and practitioners can develop effective strategies and techniques to address gender bias in NLP. The above-mentioned NLP tasks can be represented and examined using the terms specified in Table 1.

2. Transformers and gender bias

In recent years, the popularity of Language Modeling has significantly increased, mainly due to the development of transformers such as BERT, GPT-2, and XLM (Wolf et al., 2019; Gillioz et al., 2020; Braşoveanu and Andonie, 2020). These deep learning models employ a self-attention process that allows them to weigh the importance of different input data components differently (Vaswani et al., 2017). Additionally, transformer-based models can be fine-tuned for a specific downstream task, making them highly versatile. Fine-tuning requires much less data than training a language model from scratch, making these models highly efficient. However, despite their efficiency, gender bias has been observed in transformers, indicating that the problem of bias in NLP is still prevalent. Evidence of gender bias in transformer models like GPT2 (Budzianowski and Vulić, 2019), GPT3 (Floridi and Chiriatti, 2020), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) was proven by researchers across the globe through a series of experiments. To identify occupational gender bias in GPT-2, examine how the prejudice evolves with various model sizes, and contrast this bias with bias in our culture, Bolukbasi et al. (2016) conducted

Table 1
Examples of Gender Bias in different tasks.

Task	Example of Representation Bias in the Context of Gender	A1	A2	A3	References
Word Embeddings	Automatic Generation of analogies like "***Man: Woman :: Programmer: Homemaker***"	✓	✓	✓	Amazon (Anon, 2023), Caliskan et al. (2022)
Machine Translation	"He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor."	×	✓	×	Savoldi et al. (2021)
Caption Generation	An image captioning model incorrectly predicts the agent to be male	×	✓	×	Hendricks et al. (2018), Hirota et al. (2022)
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing Male noun phrases	×	✓	×	Park et al. (2018), Asyrofi et al. (2022), Kiritchenko et al. (Kiritchenko and Mohammad, 2018a)
Language Model	"He is a doctor" has a higher conditional likelihood than "She is a doctor"	×	✓	✓	Lu et al. (2020), Bordia et al. (Bordia and Bowman, 2019)

several tests. The experiments have shown that occupations became more gender-neutral as the number of trained parameters increased. According to societal statistics and all four GPT-2 models, there is a trend towards increased male bias as job salaries rise. That is, the more senior the job and the bigger its monetary compensation, the more likely it is that a man is holding that position, according to GPT-2.

Similarly, a study conducted by Brown et al. (2020) has also proven that GPT-3 contains racial and gender bias. On performing the occupational experiment, results have shown that GPT-3 found that 83% of the 388 evaluated vocations were more likely to be linked to a male identity. Also, higher-level occupations with a predominance of men included banker and professor emeritus. Another well-known AI chatbot, ChatGPT (Lund and Wang, 2023), has been accused of gender bias (Borji, 2023; Ortega-Martín et al., 2023). Kieran Snyder, a co-founder of Textio, points out that ChatGPT can start incorporating gendered presumptions into feedback that is otherwise so generic with very little effort. The feedback that is of a high caliber concentrates on an individual's work rather than their personality. It offers precise, pertinent examples. It is concise, pertinent, and straightforward. Men were noticeably more likely than women to be labeled as ambitious and confident in Textio's groundbreaking analysis of performance evaluation received by 25,000+ workers at 250+ organizations, while women were more likely to be regarded as collaborative, helpful, and outspoken. They represent the precise bias trends that appear in ChatGPT's written comments. Researchers are currently working on extending transformers to other types of input, particularly visual inputs because they have shown to be so helpful in NLP. The attention processes that the transformer employs are responsible for its performance in each of these domains. Models can selectively focus on a few pertinent parameters while ignoring others thanks to attention processes.

Also, a transformer considers (possibly) all data at once. Therefore, the influence is all against all rather than one against the next, whereas in standard neural networks, the processing of one item influences the processing through recursion and changes the way the following one is processed. Contrary to popular belief, this strategy uses less computing power. When more data spaces are processed and more become available, the use of transformers will spread and result in acceleration. Several main similarities between how people process information and learn across a wide range of tasks appear to have been incorporated into the transformer (Khan et al., 2022). The similarities to human learning provide a positive outlook on the transformer's potential in the future. In other AI disciplines, current research points to various novel applications for transformers, such as teaching robots to recognize human body movements (Jangir et al., 2022), teaching computers to understand emotions in speech, and identifying stress levels in electrocardiograms. Due to their versatility, transformers can be considered the future of AI (Han et al., 2021), and it is highly considered essential to identify and mitigate gender bias in these models.

Hence, Gender bias in Transformers is a critical issue that demands a thorough investigation, driven by its **far-reaching implications for society, ethics**, and **the advancement of artificial intelligence**. Ethically, it perpetuates and exacerbates existing societal biases, fostering unfair treatment, and discrimination, and reinforcing harmful stereotypes. This bias can potentially marginalize individuals based on their gender, deepening social inequalities and injustices. Moreover, it directly impacts users of AI systems powered by Transformers, including chatbots, virtual assistants, and language translation services, as it can generate biased or offensive responses, adversely affecting user experiences. With the increasing push towards ethical AI, understanding and addressing gender biases within such influential models is crucial in ensuring that the benefits of AI are equitably distributed. Beyond that, these biased AI systems can perpetuate gender stereotypes, associating women with caregiving roles and men with technical expertise, further entrenching societal prejudices. Gender bias also infiltrates critical areas such as hiring, admissions processes, and healthcare, potentially leading to unfair discrimination and disparities. The influence extends to language and culture, as gender bias in language models can harm the representation of gender diversity and non-binary identities, hindering linguistic inclusivity. Legal and regulatory concerns surround AI bias, with countries contemplating legislation to regulate these technologies.

Understanding gender bias in Transformers is crucial for compliance and accountability in AI development. Transformers' ability to amplify biases from their training data makes it imperative to comprehend the mechanisms behind this bias amplification, especially concerning gender bias. Moreover, gender bias often intersects with other biases, such as racial or ethnic bias, compounding its complexity and harm. Lastly, given that Transformers and similar models are at the forefront of AI research, addressing gender bias in these models is foundational for fostering more inclusive and equitable AI technologies in the future. In essence, gender bias in Transformers transcends the technical realm, posing multifaceted challenges with profound societal, ethical, and practical consequences. A comprehensive examination and mitigation of this bias are essential to ensure that AI technologies contribute positively to society, promoting fairness, equity, and inclusivity. The following are the major highlights of the survey presented in this research.

- In our research, we emphasize various scenarios of gender bias, identify sources of bias, and articulate the linguistic consequences of gender bias in the latest Transformer models. We have categorized these consequences in a clear and concise manner, enabling a comprehensive understanding of the issue.
- Furthermore, we provide a detailed overview of cutting-edge techniques utilized to detect and alleviate gender bias in Transformer Models. By thoroughly examining and analyzing each proposed methodology from diverse researchers worldwide, we also discuss the experiments conducted and datasets used to obtain the results.

Table 2
Tabulated overview of Bias Sources in NLP.

Source of bias	Description	Key characteristics	References
Data Bias	Bias present in the training data used for NLP models	Limited demographic representation, skewed patterns, and associations, temporal bias from outdated sources	Garimella et al. (2019)
Annotation Bias	Bias introduced during the process of data labeling and annotation	Assumptions and stereotypes of annotators, alignment with biases in training data, inconsistencies and disagreements among annotators	Lingren et al. (2014)
Bias from Input Representations	Bias arising from the representativeness and pre-processing of input data	Lack of diversity in input data, loss of cultural nuances and biases introduced during pre-processing steps, biases in word embeddings and contextual representations	Peng et al. (2019)
Model Bias	Bias amplified and reinforced by the NLP models themselves	Bias overamplification, reliance on discriminatory features, choice of loss objectives favoring accuracy over fairness	Kiritchenko et al. (Kiritchenko and Mohammad, 2018b), Hovy et al. (2020)

- Our survey of gender bias in Transformer models stands out as the most comprehensive to date, as we not only highlight the shortcomings of the proposed solutions mentioned above and examine future development prospects but also establish rules and ethical guidelines for measuring gender bias.

3. Why does gender bias occur?

According to a study by Hovy and Prabhumoye (2021), there can be four primary sources of bias in NLP. These include bias from data, annotations, input representations, and models, which can be summarized in Table 2.

3.1. Data bias

Biases in language training data have been extensively documented and recognized as a significant challenge in NLP systems. The prevalent sources of training data for NLP models often come from well-known news outlets, which tend to represent a limited demographic profile. These sources predominantly reflect the perspectives of individuals who are white, upper-middle-class, middle-aged, and educated (Garimella et al., 2019). Consequently, NLP models trained on such data inherit and perpetuate the demographic bias in the training samples, leading to biased predictions and outputs. The biased behavior of NLP models stems from learning patterns and associations from the training data. If the data is skewed towards a specific demographic, the models will reflect and reinforce that bias. For example, a model trained on news articles that predominantly feature male politicians may associate leadership roles with masculinity, leading to biased predictions in gender-related tasks. Furthermore, many syntactic tools, such as taggers and parsers, rely on outdated newswire data from the 1980s and 1990s. These tools can inadvertently reinforce biases by assuming that everyone speaks and writes in a manner similar to journalists from that era. This temporal bias can result in models failing to understand or accurately represent the language used by diverse demographic groups, further perpetuating linguistic inequality. Given these challenges, it is crucial to consider the demographic representation within the chosen text data collection. Even seemingly neutral or unbiased datasets can carry latent biases due to the inherent demographic signals embedded in language itself. Therefore, researchers and practitioners must be mindful of their training data's limitations and potential biases. By carefully considering the demographic groupings represented in the data and actively working towards mitigating biases, NLP researchers and practitioners can strive to develop more fair and equitable systems. This entails promoting diversity and inclusivity in both the training data and the development process, leading to NLP models that better understand and respect the rich linguistic variations present in society.

3.2. Annotation bias

Annotation bias, also known as **label bias**, is a phenomenon in NLP where human annotators inadvertently introduce biases into labeled data used for training and evaluating NLP models (Lingren et al., 2014). This bias can significantly impact the performance and fairness of these models. There are several factors that can contribute to annotation bias. Annotators may hold certain assumptions or stereotypes about the language or task they are working on, which can influence their labeling decisions. For example, they might have preconceived notions about the sentiment of a particular text or the gender associated with certain occupations. Furthermore, annotators can be influenced by the data they are annotating, especially if the data itself contains biases, as mentioned in the previous section. If the training data exhibits imbalances or reflects societal biases, annotators may inadvertently align their labels with those biases, reinforcing them in the labeled dataset. Addressing annotation bias is essential for developing NLP models that are fair, robust, and respectful of diverse perspectives. It is essential to be aware of its presence and take proactive steps during the annotation process to mitigate annotation bias. Providing clear guidelines to annotators that explicitly address potential biases and instruct them to label based on the content rather than their assumptions or stereotypes can help reduce bias. Regular training and discussions with annotators can promote awareness and sensitivity towards bias. Using multiple annotators and measuring inter-annotator agreement can help identify and address inconsistencies or biases in the labeled data. Adjudication or consensus mechanisms can resolve disagreements among annotators, ensuring a more balanced and unbiased representation of the data. Furthermore, ongoing efforts are being made to develop methods that explicitly model and mitigate annotation bias while training NLP models. By explicitly accounting for the biases introduced during annotation, reducing their impact on the model's predictions and improving fairness is possible. By acknowledging and actively mitigating annotation bias, the NLP community can strive towards more accurate and equitable NLP systems that better reflect natural language's varied nuances and characteristics.

3.3. Bias from input representations

Bias from input representations in NLP refers to the introduction of bias into the input data used for training and evaluating NLP models. These biases, often referred to as semantic biases, can arise from various sources and have significant implications for the fairness and accuracy of NLP systems. One common source of bias is the lack of representativeness in the input data. If the training data does not adequately reflect the diversity of the population or the specific task it aims to address, biases can emerge. For example, if the training data predominantly represents a particular demographic group or region, the model's predictions may be skewed towards that group's perspectives and experiences, leading to biased outputs. Another factor contributing to bias in input representations is the pre-processing of data. Pre-processing steps, such as text normalization, tokenization, or stemming, can inadvertently introduce biases. For example, certain linguistic variations or expressions commonly used by specific demographic groups may be overlooked or normalized, resulting in a loss of cultural nuances and potential bias in the representations.

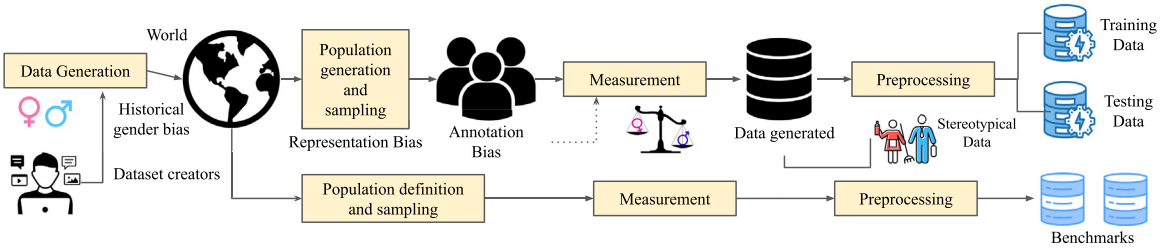


Fig. 4. Gender Bias induced from Data Generation.

Studies have demonstrated that word embeddings can detect racial and gender biases, even in well-labeled and balanced datasets. These biases can arise due to societal prejudices reflected in the training data, leading to biased predictions and outputs from NLP models. Contextual representations learned by large pre-trained language models, such as BERT and GPT, are also susceptible to biases. These models are typically trained on vast amounts of internet text, including societal biases in online content. Consequently, these models can replicate and perpetuate biases, often mirroring societal biases. Numerous studies have documented and quantified biases in NLP models and their input representations, highlighting the importance of addressing these issues. Recognizing and understanding the biases generated during the data generation process is a critical step towards mitigating them. Addressing bias in input representations requires a multi-faceted approach. It involves diversifying training data sources to ensure the representation of various demographic groups and perspectives. Regularly evaluating and auditing the models for biases and developing debiasing techniques are crucial to mitigating these biases. The above-mentioned biases are collectively represented as the biases generated in the process of data generation, which can be depicted in Fig. 4

3.4. Model bias

Languages are dynamic and constantly evolving, making capturing their complexity and nuances challenging even with a large dataset. Using a small subset of data can only provide a limited and temporary snapshot of language, which is why relying solely on "better" training data is not a comprehensive solution to address bias in NLP models. Furthermore, machine learning models tend to amplify the behaviors and patterns they are exposed to, including biases present in the training data. Studies such as Kiritchenko and Mohammad (2018b), Hovy et al. (2020) have explored the compounding effect of bias in newer models, highlighting the phenomenon known as bias overamplification. This refers to the tendency of machine learning models to disproportionately amplify and reinforce biases rather than mitigate them. One contributing factor to bias overamplification in language models is the choice of loss objective used during training. Often, these objectives prioritize improving the model's prediction accuracy, which can incentivize the model to exploit spurious correlations or irregularities in the training data. As a result, the model may rely on certain discriminatory features, such as gender or race, to achieve higher accuracy, even if those features are irrelevant to the task. This behavior is challenging to detect until a consistent pattern of bias is identified and examined. Addressing bias overamplification requires a more nuanced and comprehensive approach beyond improving the training data. It involves reevaluating the loss objectives and training methods to incorporate fairness and mitigate biases. Researchers and practitioners are actively exploring techniques to promote fairness and reduce bias in NLP models, such as incorporating fairness constraints, developing debiasing algorithms, or redefining evaluation metrics to account for bias. Moreover, addressing bias overamplification requires collaboration and engagement with diverse stakeholders, including linguists, ethicists, and impacted communities. These collaborations can help in uncovering and understand the complexities of bias in language models and develop more holistic approaches to mitigate bias overamplification. Fig. 5 gives a clear pictorial representation of model bias.

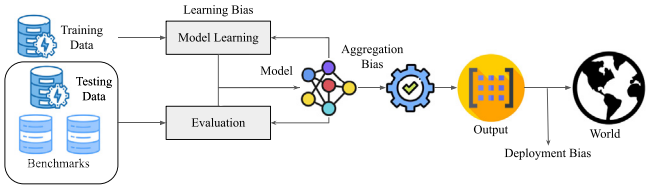


Fig. 5. Gender Bias Induced from Model Building.

4. Examining gender bias detection techniques in transformer models

When evaluating gender bias in transformer models, it is important to use a combination of metrics to comprehensively understand the model's performance. While individual metrics can provide insights into specific aspects of bias, they may not capture the full extent of gender bias present in the model. As a result, combining multiple metrics can provide a more accurate assessment of gender bias. There exist a variety of metrics that can be employed to evaluate gender bias in transformer models whose summary can be illustrated in Table 3

4.1. WEAT score

The **WEAT (Word Embedding Association Test) score** measures the degree of association between two sets of words based on their embedding vectors in a language model, such as a transformer model. Specifically, it measures the degree of association between a set of target words and a set of attribute words. In the context of gender bias in transformer models, the WEAT score can be used to evaluate whether the model is exhibiting biased associations between gendered words and certain attributes (such as career vs. family). A higher WEAT score indicates a stronger association between the target and attribute sets, which could suggest the presence of bias in the model. The WEAT score is based on the concept of "cosine similarity", which is a measure of the similarity between two vectors. In the context of word embeddings, cosine similarity is used to measure the similarity between the vectors representing two words. To calculate the WEAT score, four sets of words are first defined: *Set A* consisting of words that are stereotypically associated with one gender *"man", "male", "he", "brother"*, *Set B*, consisting of words that are stereotypically associated with the other gender *"woman", "female", "she", "sister"*. *Set X* consists of words that are associated with a specific attribute *"career", "professional", "executive"* and *Set Y* consists of words that are associated with a contrasting attribute *"family", "home", "parent"*. The WEAT score is then calculated as the difference between the average cosine similarities of words in sets A and X, and the average cosine similarities of words in sets B and X, normalized by the standard deviation of the cosine similarities of all words in the sets:

$$\mathrm{W}(X, Y, A, B) = \frac{\sum_{w \in X} \cos(w, A) - \sum_{w \in X} \cos(w, B)}{\sqrt{\sum_{w \in X} (\cos(w, A) - \cos(w, B))^2}} \quad (1)$$

where $cos(w, S)$ is the cosine similarity between the word w and the average of the vectors in set S. A higher absolute value of the WEAT

Table 3
Tabulated overview of Gender Bias Detection Techniques in Transformer Models.

Technique	Methodology	Limitations
WEAT score	1. Measures the degree of association between gendered words and certain attributes using word embedding vectors. 2. Higher score indicates stronger association and potential bias.	1. Word Embedding limitations and a limited set of words may not capture the full range of gendered associations. 2. Lack of Contextual Information and Subjectivity of Attribute Definitions.
Equalized Odds	1. Evaluates whether a model's predictions are equal across different demographic groups. 2. Specifically, compares true positive rates and false positive rates for males and females.	1. Binary Classification Focus may not directly apply to more nuanced gender categories or non-binary gender identities. 2. Assumption of Independence and Trade-offs and Information Loss resulting in reduced overall accuracy
Counterfactual evaluation	1. Measures gender bias by assessing the impact of gender swapping on model performance. 2. Compares accuracy on original and gender-swapped test sets to quantify the degree of bias.	1. Availability of Counterfactual Data and the definition of Counterfactual Scenarios; selecting unrealistic or arbitrary counterfactual scenarios can lead to misleading assessments of bias 2. Limited Scope of Counterfactuals and Causal Inference Challenges exist in this methodology
BLEU Score	1. Calculates the similarity between the model's output and the original sentence using gender-swapped sentence versions. 2. Used to estimate gender bias based on score differences.	1. BLEU primarily measures the lexical overlap between the generated and reference translations. It does not capture language's semantic or contextual aspects, including gender bias. 2. BLEU does not explicitly consider gender-related biases in the translations. It treats all words and phrases equally without accounting for potential stereotypes, imbalances, or unequal treatment based on gender.
Stereoset	1. Utilizes crowd-sourcing to evaluate stereotyped evaluations made by MLMs on gender and career. 2. Provides bias scores on a scale of 0 to 100, with lower scores indicating less bias.	1. Limited to MLMs as Stereoset relies on predefined sets of stereotypical sentences, which may not cover the entire spectrum of possible biases.
Attention Maps	1. Analyzes attention maps in transformer-based models to detect gender bias. 2. Compares the relation degree between genders and occupations based on attention scores to identify bias-contributing modules.	1. Attention scores themselves do not always have a direct and intuitive correspondence to the importance or significance of specific features or connections. Interpreting attention maps requires careful analysis and domain expertise. 2. Transformer models often exhibit sparse attention, meaning that only a small subset of the input tokens receive significant attention weights. This sparsity can limit the granularity of the analysis and make it difficult to capture fine-grained biases.

score indicates a stronger association between the attributes X and Y and the gendered words in sets A and B. **A WEAT score of 0 indicates that there is no difference in the association between the two sets of gendered words and the attributes X and Y.** This metric was illustrated by Silva et al. (2021) with the investigation of gender bias in transformers like GPT-2 and XLNet by employing WEAT as one of the methodologies used to detect gender bias. According to the findings, RoBERTa is one of the most consistently biased transformers among other models.

4.2. Equalized odds

Equalized Odds is a fairness metric that can be used to detect gender bias in transformer models (Awasthi et al., 2020; Garg et al., 2020). The metric measures the degree to which a model's predictions are equal across different demographic groups, such as males and females. In the context of gender bias, the metric can be used to assess whether the model is making equally accurate predictions for male and female inputs. The Equalized Odds metric is calculated by comparing the **True Positive Rates (TPRs)** and **False Positive Rates (FPRs)** across different demographic groups for a given prediction task. In the context of gender bias, this typically involves comparing the TPRs and FPRs for male and female inputs. The TPR measures the proportion of positive cases that are correctly identified by the model, while the FPR measures the proportion of negative cases that are incorrectly identified as positive by the model. By comparing the TPRs and FPRs across demographic groups, the Equalized Odds metric can provide a measure of how fairly the model is making predictions for different groups. A lower Equalized Odds score indicates a lower level of gender bias in the model's predictions, as it suggests that the model is making equally accurate predictions for males and females. The TPRs and FPRs for each demographic group are first computed to calculate the Equalized Odds metric. The metric is then calculated as the maximum difference in TPRs or FPRs across demographic groups. Specifically, the metric is the maximum of the absolute differences in TPRs or FPRs for any given threshold:

$$\text{Equalized Odds} = \max_{t\in[0,1]} |\text{TPRmale}(t) - \text{TPRfemale}(t)| \tag{2}$$

$$\text{Equalized Odds} = \max_{t\in[0,1]} |\text{FPRmale}(t) - \text{FPRfemale}(t)| \tag{3}$$

4.3. Counterfactual evaluation

Counterfactual evaluation is a technique used to measure gender bias in transformer models by assessing the impact of gender swapping on model performance. The method involves modifying the gender of words in a dataset and observing the effect on the model's accuracy and other performance metrics. To apply this method, a dataset is first split into a training set and a test set. Then, the gender of the words in the test set is modified by swapping gendered pronouns or replacing gendered words with gender-neutral alternatives. For example, the word "he" could be replaced with "they", or the name "John" could be replaced with "Alex". The modified test set is then used to evaluate the model's performance, comparing the results to those obtained from the original test set. The difference in accuracy and other performance metrics between the two sets is used to calculate the degree of gender bias present in the model. Let the original test set be denoted by X and its associated labels by Y. Let the modified test set, where gendered words have been replaced with gender-neutral alternatives, be denoted by X', with associated labels Y'. Let the model's predicted labels on X and X' be denoted by Y_{hat} and Y'_{hat}, respectively. The first step is to calculate the baseline performance of the model on the original test set:

$$acc_{orig} = \frac{1}{|X|}\sum_{i=1}^{|X|}[Y_i = Y_{hat,i}] \tag{4}$$

where acc_{orig} is the accuracy of the model on the original test set, $|X|$ is the number of examples in the test set, Y_i is the true label of example i, and $Y_{hat,i}$ is the predicted label of example i. Next, the performance of the model on the modified test set is calculated:

$$acc_{mod} = \frac{1}{|X'|}\sum_{i=1}^{|X'|}[Y'_i = Y'_{hat,i}] \quad (5)$$

where acc_{mod} is the accuracy of the model on the modified test set, $|X'|$ is the number of examples in the modified test set, Y'_i is the true label of the gender-swapped example i, and $Y'_{hat,i}$ is the predicted label of the gender-swapped example I. The degree of gender bias in the model can be calculated as the difference between the two accuracies:

$$bias_{gender} = acc_{orig} - acc_{mod} \quad (6)$$

A positive value of $bias_{gender}$ indicates that the model is biased towards the original gender, while a negative value indicates a bias towards the opposite gender. A value of zero indicates no gender bias in the model. The counterfactual evaluation method provides a way to measure gender bias in transformer models without relying on external benchmarks or human annotations. By simulating the impact of gender-swapping on model performance, the technique can identify cases where the model relies on gendered cues to make predictions rather than on other relevant information in the text. One limitation of the counterfactual evaluation method is that it only measures the impact of gender on performance in a binary sense (i.e., male versus female). Other factors, such as race, ethnicity, or sexuality, may also contribute to bias in the model but are not captured by this method. Additionally, the technique assumes that gender-neutral replacements are available for all gendered words in the dataset, which may not always be the case. Overall, counterfactual evaluation is a valuable method for detecting and quantifying gender bias in transformer models, providing a complement to other techniques such as WEAT and Equalized Odds.

4.4. BLEU score

BLEU (Bilingual Evaluation Understudy) score is a metric commonly used to evaluate the performance of machine translation systems (Wołk and Marasek, 2015; Papineni et al., 2002). However, it can also estimate gender bias in transformer models. The basic idea is to use gender-swapped versions of sentences as inputs and compare the similarity of the model's outputs with the original sentences. To apply this method, a dataset is first split into a training set and a test set. Then, gender-swapped versions of sentences in the test set are created, where the gendered words are replaced with their gender-neutral counterparts. The BLEU score measures the degree of overlap between the n-grams (subsequences of n words) in the model's output and the original sentence, with higher scores indicating greater similarity. The BLEU score is calculated using the following equation:

$$BLEU = BP \times exp(\frac{1}{n}\sum_{i=1}^{n} log p_i) \quad (7)$$

where BP is the brevity penalty, which adjusts the score based on the length of the output sentence compared to the original sentence, and p_i is the precision score for n-grams of length i. The precision score is calculated as the number of n-grams in the model's output that appear in the original sentence divided by the total number of n-grams in the model's output. To estimate gender bias using the BLEU score, the average score for male and female gender-swapped sentences is calculated and compared. If the model consistently produces higher BLEU scores for male gender-swapped sentences than for female ones, this indicates a bias towards male language. Conversely, if the model consistently produces higher BLEU scores for female gender-swapped sentences than for male ones, this indicates a bias towards female language. One advantage of the BLEU score is that it is a widely used and standardized metric, making comparing results across different studies easier (see Fig. 6).

4.5. Stereoset

Robinson (2021) proposed a technique for identifying gender bias in both traditional and Medical **Masked Language Models (MLMs)** such as SciBERT (Beltagy et al., 2019) and BioClinicalBERT. Medical MLMs refer to language models that are specifically pre-trained on medical text. These models have the potential to improve the accuracy and speed of medical text analysis and provide new insights into clinical data. Unlike general-purpose MLMs, medical MLMs are pre-trained on a large corpus of medical text, including scientific publications, clinical notes, and electronic health records. This allows the models to capture the unique language and terminology used in the medical field and the specific contexts in which these terms are used. The proposed methodology, called StereoSet, uses a collection of 17,000 test words and crowd-sourcing to quantify stereotyped evaluations about gender and career made by MLMs. StereoSet evaluates the most likely word chosen by an MLM to fill in the blank in intra-sentence and inter-sentence Context Association Tests (CATs). Bias scores are given on a scale of 0 (strong bias) to 100 (extremely low or no bias), and BERT achieved a gender bias score of 63, while RoBERTa achieved a score of 73. On the other hand, the medical MLMs exhibited more bias in all categories than the general-purpose MLMs, except for SciBERT, which showed a better race bias score of 55 than BERT's 53. The medical MLMs also showed more gender and religious biases compared to the general-purpose MLMs. The evaluation of four medical MLMs for stereotyped assessments about race, gender, religion, and profession revealed lower performance compared to general-purpose MLMs. These medically-focused MLMs differ considerably in their training source data, which likely contributes to the differences in the ratings for stereotyped biases from the StereoSet tool. Overall, this study highlights the importance of considering and addressing biases in NLP systems, particularly those used in sensitive areas such as healthcare and science, where accurate and unbiased results are crucial.

4.6. Attention maps

Li et al. (2021) presented a novel gender bias detection method for transformer-based models by utilizing attention maps. The authors propose an intuitive gender bias judgment method by comparing the relation degree between genders and occupations based on attention scores. They also design a gender bias detector by modifying the attention module and inserting it into different positions of the model to present the internal gender bias flow. By scanning the entire Wikipedia, a BERT pre-training dataset, the authors draw a consistent gender bias conclusion. Their findings show that attention matrices, W_q and W_k, introduce much more gender bias than other modules, including the embedding layer. The bias degree changes periodically inside the model, where the attention matrix Q, K, V, and the remaining part of the attention layer enhance gender bias, while the averaged attentions reduce the bias. This study is the first attempt to investigate gender bias inside transformer-based models, using BERT as an example, and provides insights into the mechanisms that contribute to gender bias in NLP models.

4.7. Discussion

The interrelation and distinction between various metrics measuring biases and fairness in NLP can be understood through straightforward examples. Take WEAT (Word Embedding Association Test), which measures the association between sets of target words and attribute words. If one were to evaluate Male names John, Mike versus Female names Mary, Susan against Career job, salary and Family home, children attributes, a positive WEAT score might imply a stronger association of male names with career-related words. Then there is StereoSet, which gauges biases in sentence predictions. In a scenario where the statement "John is a " gets completed as a "software engineer" while "Mary

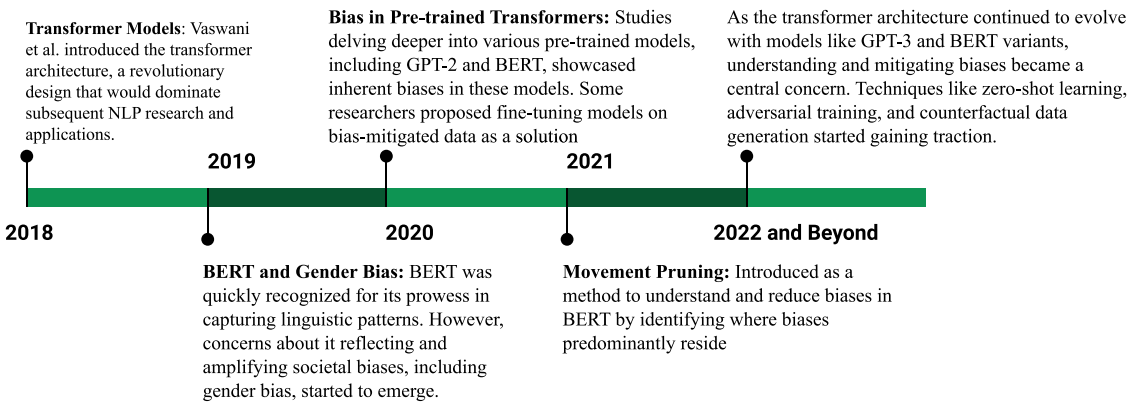


Fig. 6. Literature Timeline of Gender Bias in Transformer Models.

is a " gets completed as "nurse", it clearly reflects an inherent bias. Equalized Odds is another metric, ensuring consistent false positive and true positive rates across demographic groups. For instance, if 20% of deserving males and 5% of deserving females get wrongly rejected for a loan, the odds are not balanced. Counterfactual Evaluation then inspects how prediction outputs shift when specific input attributes, like gender, change. A sentence like "He is a doctor" might have a drastically different sentiment score or output when flipped to "She is a doctor", which would imply bias. Finally, the BLEU Score, usually used to gauge translation quality, can be adapted to this context. Given an original sentence, "He is a good teacher", and its gender-swapped counterpart, "She is a good teacher" if a model translates the original as "He is a skilled instructor", the BLEU score would then assesses how closely the model's translation aligns with the gender-swapped version. High BLEU scores would suggest unbiased treatment of the gender-swapped sentences. It is essential to note that these examples are a simplification, and real-world applications of these metrics can be more intricate.

Gender bias in transformer models can be meticulously evaluated using these spectrum of metrics, each offering a unique viewpoint. For instance, the WEAT score and StereoSet delve into word embeddings or masked language model predictions to uncover implicit biases by examining associations between words or categories. On the other hand, Equalized Odds and Counterfactual evaluation provide insights into the model's prediction disparities among different demographic groups, offering a performance-level view of biases. BLEU score, emphasizing output quality, measures the likeness between model outputs and gender-swapped input sentences. In terms of granularity, while WEAT and StereoSet can unearth more nuanced biases in the model's latent space, Equalized Odds, Counterfactual evaluation, and BLEU score provide a direct lens on explicit biases. StereoSet's design allows it to be adjusted to specific domains, such as medicine, to detect biases in specialized contexts. Differing in their measurement techniques, WEAT employs cosine similarity in embedding space, an unsupervised measure, whereas Equalized Odds leans on true positive and false positive rates, requiring ground truth labels. Counterfactual evaluation, resembling BLEU's gender-swapping method, focuses on performance over output similarity. Notably, while Counterfactual evaluation and BLEU often operate within a binary gender framework, StereoSet encompasses a broader net, capturing biases related to factors like race and religion. Understanding the interplay between these metrics enables the development of holistic strategies for bias intervention. For instance, while metrics like Equalized Odds guide prediction task interventions, others like WEAT can directly inform efforts to debias word representations. In essence, this diverse toolkit of metrics, when used in tandem, offers a comprehensive assessment of gender bias, revealing both its origins and potential remedies.

In a comparative analysis between various gender bias detection techniques in transformer models, distinct contrasts emerge. Techniques such as Counterfactual evaluation and Stereoset provide direct insights into bias, furnishing straightforward evidence of prejudice. On the other hand, methods like Attention Maps serve a more diagnostic purpose, offering a lens into the model's internal mechanics rather than clear-cut biases. In terms of breadth and depth, while Stereoset's predefined sentences might fall short of capturing the full spectrum of potential biases, Attention Maps compensate with their granular insights, highlighting token-level interactions within the model. For those seeking comprehensive gender bias detection, it is advisable to commence with quick evaluations using methods like the WEAT score or Stereoset. Following these preliminary checks, a more profound analysis can be undertaken by amalgamating direct techniques like Counterfactual evaluation with diagnostic ones, exemplified by Attention Maps, ensuring surface-level and in-depth scrutiny of biases. The above conclusion can be aligned with the following strategies to mitigate gender bias.

5. Overview of gender bias estimation techniques in transformer models

Over the years, researchers have proposed various methodologies for Gender Bias mitigation in transformer models, as depicted in the literature timeline. These methods include altering the training data, adjusting model architecture, and incorporating additional constraints during training. Each of these approaches has its own advantages and limitations. This section provides an overview of the different methodologies used for Gender Bias mitigation among various transformer models with its summary being portrayed in Table 4. We discuss the strengths and weaknesses of each approach and highlight the current state-of-the-art methods in this field. Through this discussion, we hope to provide a comprehensive understanding of Gender Bias mitigation in transformer models and its importance in creating fair and equitable NLP systems.

5.1. Movement pruning

A technique used to inspect the gender bias in pre-trained language models using the attention layers was proposed by Joniak and Aizawa (2022), also known as movement pruning. When some weights are disabled or removed from a neural network, the process is known as pruning. The authors modified movement pruning, allowing one to select a low-bias subset of a given model or, more precisely, to identify the model weights whose removal causes an arbitrary debiasing objective to converge. The proposed strategy is innovative since it combines debiasing, weight freezing, and movement pruning. It also investigates if gender bias exists in a BERT model and suggests ways to improve an existing debiasing technique. Gender bias was used by the authors to demonstrate how to use their framework, and they discovered that the bias is primarily stored in the intermediate layers of BERT. The approach uses movement pruning to identify a subset that has less

Table 4
Literature review involving the reduction of Gender Bias in Transformer Models.

Author	Strengths	Weaknesses	Notable findings
Joniak et al. (Joniak and Aizawa, 2022)	1. Identifies weights that, when removed, lead to convergence of debiasing objectives. 2. Combination of debiasing, weight freezing, and movement pruning. 3. Enables model optimization by removing bias-related weights.	1. Focuses on weight removal without examining other aspects of bias. 2. May affect model performance.	1. Gender bias is primarily stored in intermediate layers of BERT 2. A direct relationship between model bias and performance.
Bao et al. (Bao and Qiao, 2019)	1. Utilizes knowledge learned from pre-existing models to improve performance on a new task. 2. Reduces the need for extensive training data. 3. Enhances performance on tasks requiring less data.	1. Relies on the availability of pre-existing models. 2. May not effectively address task-specific biases. 3. Requires careful fine-tuning to balance general knowledge and task-specific features.	1. Improvement in coreference system performance on the GAP dataset. 2. Effective reuse of pre-trained BERT knowledge for improved task performance.
Vig et al. (2020)	1. Investigates the mechanisms influencing gender bias in models. 2. Identifies specialized model components responsible for bias. 3. Examines effects flowing directly and indirectly through mediators.	1. Limited to specific models (GPT2, XLNet, RoBERTa, DistilBERT). 2. Does not provide direct mitigation strategies. 3. Requires careful analysis and interpretation of results.	1. Gender bias effects concentrated in specific model components. 2. Highly specialized behavior observed in certain model components.
Bhardwaj et al. (2021)	1. Evaluates gender bias in CLMs using regression models. 2. Identifies gender-specific word dependencies. 3. Identifies gender subspace in word embeddings.	1. Focuses on word embeddings and may not capture all aspects of bias. 2. Does not provide direct mitigation strategies.	1. CLM predictions are significantly influenced by gender-specific words. 2. Identification of gender subspace and specific gender directions in BERT.
Basta et al. (2020)	1. Incorporates prior phrase and speaker information in NMT models. 2. Improves translation quality and reduces gender bias. 3. Efficient architecture with reduced training parameters.	1. Limited to decoder-based NMT models. 2. Relies on the availability of previous sentence information. 3. May not fully address biases in the source text.	1. Improved translation quality and reduced gender bias in NMT models. 2. Robustness achieved by including previous sentence information.
Bartl et al. (2020)	1. Mitigates bias by fine-tuning BERT on a modified corpus	1. Effectiveness may vary across languages with complex gender systems. 2. It does not work well for languages like German, which have a sophisticated morphology.	1. Successful bias reduction in English but challenges in German due to language characteristics.
Kaneko et al. (2022)	1. Evaluates bias across languages using English attribute word lists and parallel corpora.	1. Requires parallel corpora between the target language and English	1. Identified gender-related stereotypes in MLMs across multiple languages
De et al. (de Vassimon Manela et al., 2021)	1. Introduces measures to analyze and reduce gender bias in contextual language models	1. Trade-off between skew and stereotype in out-of-the-box models.	1. Optimized models significantly reduce both skew and stereotype.

bias than the original model given a model and a debiasing aim. If a model produces more invariance, it may become faster and smaller while preserving its previous performance. They also noticed that there is a direct relationship between model bias and performance.

5.2. Transfer learning

Transfer learning is a machine learning technique in which a model created for one task is utilized as the foundation for a model on another. In other words, it involves leveraging the knowledge learned from a pre-existing model to improve the performance of a new model on a related or different task. The main advantage of transfer learning is that it can significantly reduce the amount of training data and time required to achieve high performance on a new task. The idea is to use this model as a starting point and then fine-tune it on the new task by updating the weights of some of the layers or adding new layers to the model. This way, the model can learn task-specific features while retaining the general knowledge learned from the pre-existing model. Mathematically, transfer learning can be represented as follows. Let D_s and D_t be the source and target domains, respectively, where D_s is the domain on which the pre-existing model is trained and D_t is the domain of the new task. Let f_s be the pre-existing model and f_t be the new model. The goal of transfer learning is to learn a mapping

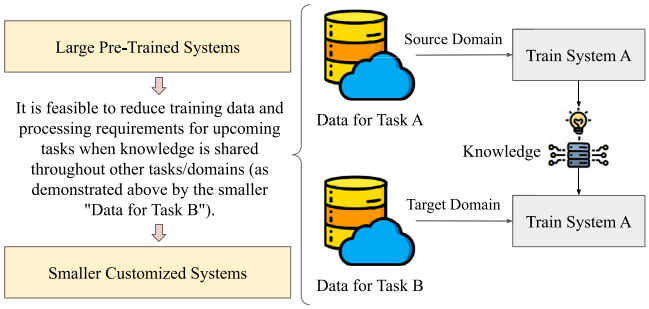


Fig. 7. Transfer Learning Paradigm: Knowledge sharing between domains.

$h : D_s \rightarrow D_t$ that transfers knowledge from f_s to f_t. The conceptual overview of transfer learning can be depicted in Fig. 7

Bao and Qiao (2019) investigated transfer learning from pre-trained models to improve task performance with little data. It has been demonstrated that the majority of current representative coreference systems suffered on the **GAP dataset** (Beamer et al., 2015), performing only mediocrely overall and with significant gender differences in performance. These coreference systems' uneven training datasets or the systems' architecture may be to condemn for this. To enhance

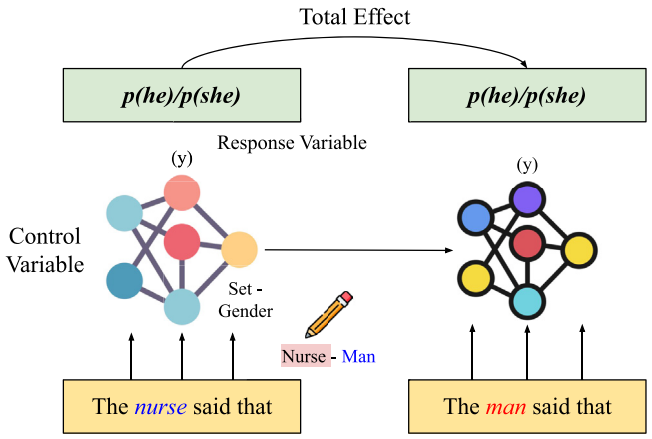


Fig. 8. Casual Mediation Analysis.

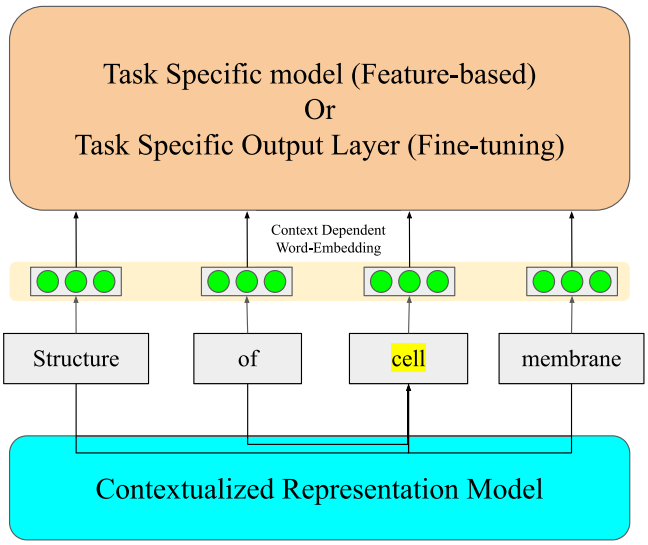


Fig. 9. Contextual Addition on Task Specific Model.

the performance of tasks requiring less data, the authors of this work investigated transfer learning from trained models. Furthermore, using data from the **Caliskan dataset** (Wolfe and Caliskan, 2021), a statistical experiment was conducted to examine gender bias in word and sentence-level embeddings. Several efficient ways to reuse pre-trained BERT knowledge in this shared work are offered and compared. The resulting system outperforms off-the-shelf resolvers significantly, with balanced prediction performance for the two genders.

5.3. Casual mediation analysis

The goal of **Causal Mediation Analysis**, also known as CMA, is to determine the extent to which intermediary factors mediate a treatment effect. CMA separates a treatment's overall impact into its direct and indirect effects. The mediator, which is defined as the intermediate variable in the casual path, transmits the indirect effect to the result. The mediation package is made to execute CMA with the sequential ignorability assumption. Using the above concept, Vig et al. (2020) proposed a method to investigate the mechanisms that allow information to move from input to output via numerous model components. The authors described a gender-specific anti-stereotypical intervention set-gender that transforms the profession **nurse** into **man**. The total effect is denoted by the change in the response variable and the methodology is illustrated in Fig. 8. They categorized the gender bias effects as sparse, concentrated in a limited area of the network, synergistic, enhanced or suppressed by different components, and decomposable into effects flowing directly from the input and indirectly through the mediators. Using three datasets designed to assess a model's susceptibility to gender bias, the authors investigated the function of individual neurons and attention heads in influencing gender bias. This mediation study demonstrates that the effects of gender prejudice are focused on specific model components that may exhibit highly specialized behavior. The transformers involved in this study are **GPT2, XLNet, RoBERTa,** and **DistilBERT**.

5.4. MLP regression

Bhardwaj et al. (2021) addressed that CLMs are prone to learning the dataset's inherent gender bias. As a result, changing gender words—such as switching "he" for "her" or using gender-neutral words—can lead to dramatically different predictions from downstream NLP models. They concentrated on a well-known CLM, **BERT**. For several NLP tasks, they trained a basic regressor using BERT's word embeddings (Reimers and Gurevych, 2019) and then evaluated the gender bias in regressors using an equality evaluation corpus. Ideally, depending on the design, the models should not accept input that contains gender-specific information. The findings show that the system's predictions significantly rely on gender-specific words and phrases. The authors also argue that eliminating gender-specific information from word embedding can reduce such biases. They consequently discovered pathways for each layer of BERT that predominantly encode gender information. The area created by these directions in the semantic space of word embeddings is known as the gender subspace. The authors also provided one primary direction for each BERT layer to detect fine-grained gender directions. This prevents other vital information from being overlooked and does away with the requirement that gender subspace is realized in several dimensions. According to experiments, such systems operate better when components are not embedded. According to experiments, removing embedding components that point in these directions significantly minimizes the bias caused by BERT in downstream tasks.

5.5. Contextual addition

Basta et al. (2020) conducted a study to determine whether recently proposed MT approaches significantly contribute to reducing biases in document-level and gender-balanced data. The authors proposed contextual addition, which is illustrated in Fig. 9, and speaker id methodology in a decoder-based NMT model (Luo, 2019). Their work examined the techniques for incorporating the prior phrase and speaker information into a decoder-based neural MT system, also named WinoMT. Their experiments' architecture only includes the decoder portion of the well-known Transformer, which minimizes training parameters and streamlines the model. The authors noted that WinoMT is a test set lacking speaker identity and information at the document level; therefore, translation using their methodology is carried out without these details. As a result, the system becomes more robust when the information from the previous sentence is included; thus, it is okay for the authors to draw inferences without it. The results show improved translation quality (+1 BLEU point) and gender bias mitigation (+5% accuracy).

5.6. Counterfactual data substitution

In addition to traditional word embeddings, evaluating biases contained in their replacements is crucial, according to Bartl et al. (2020). By examining connections between gender-denoting target words and profession names in English and German and comparing the findings to actual workforce numbers, they used BERT to quantify gender

bias. They also reduced bias by fine-tuning BERT on the GAP corpus after applying Counterfactual Data Substitution (CDS). The gender mitigation technique CDS operates at the corpus level. It operates by randomly applying an intervention to half of a corpus's documents. This intervention aims to maintain the grammatical consistency of a document while inverting all of the gendered terminologies inside it. They demonstrated that while this method of measuring bias works well for languages like English, it does not work well for languages like German, which have a sophisticated morphology and gender-marking system. Their findings highlight the importance of looking at bias and mitigation strategies across languages, especially in light of the current focus on extensive, multilingual language models.

5.7. Multilingual bilingual evaluation

Kaneko et al. (2022) highlighted that existing bias evaluation methods require stereotypical sentence pairings with the same context and attribute terms (e.g., He/She is a nurse). Without requiring manually annotated data, they presented the Multilingual Bilingual Evaluation (MBE) score for assessing bias in several languages using just English attribute word lists and parallel corpora between the target language and English. They evaluated MLMs in eight different languages using the MBE and discovered that all of them include gender-related stereotypes. They manually created datasets for gender bias in Japanese and Russian to evaluate the MBE's validity. The findings show a strong correlation between the gender bias MBE scores and those derived from those mentioned above personally created datasets and the existing English datasets.

5.8. Skew and stereotype methodology

In a recent study addressing the issue of WinoBias pronoun resolution, de Vassimon Manela et al. (2021) have proposed two measures, namely skew, and stereotype, to analyze and quantify gender bias present in contextual language models. The skew measure aims to reduce stereotypes but may result in increased skew, while the stereotype measure seeks to optimize models using a larger gender-balanced dataset, thereby minimizing both skew and stereotype. The study compared the performance of the optimized BERT model with its unaugmented, fine-tuned counterpart, demonstrating that the optimized model significantly reduces both skew and stereotype. Additionally, the researchers found that out-of-the-box models exhibit a trade-off between skew and stereotype, with RoBERTa and ALBERT-xxlarge displaying reduced skew at the cost of higher stereotypes, while DistilBERT and BERT models have high skew and low stereotypes.

5.9. Key findings and comparative analysis

In the realm of language models, several methodologies and techniques have been employed to understand and mitigate gender bias. The innovative technique of Movement Pruning harnesses a modified pruning strategy to isolate and study a subset of a given model, leading to the striking revelation that gender bias is predominantly stored in the intermediate layers of models like BERT. On the other hand, Transfer Learning, an influential approach in machine learning, capitalizes on the strengths of pre-trained models by transferring their knowledge and adapting them to new tasks. This approach, in particular, shed light on the prominent gender discrepancies in coreference resolution systems, even when applied to tasks with scarce data. Venturing into the intricacies of how biases emerge and are perpetuated, the Causal Mediation Analysis provides a nuanced understanding. Through the lens of mediators or intermediary components, it discerns that gender bias effects are neither uniformly distributed nor random; instead, they are particularly concentrated in specific areas and can be either amplified or muted by different parts of the model.

Further deepening this exploration, MLP Regression, which uses BERT's word embeddings to craft regressors, highlighted a crucial vulnerability in our systems: simple gender word substitutions can dramatically skew predictions, underscoring the embedded biases. This finding dovetails with the efforts of Contextual Addition, a technique that bolsters translation quality by incorporating information from preceding sentences, consequently leading to a marked reduction in gender bias in translations. However, while these methods offer promising strides, Counterfactual Data Substitution points to the arduous challenge of addressing biases across diverse languages, especially those with intricate morphological nuances. Echoing this cross-linguistic concern, the Multilingual Bilingual Evaluation has not only corroborated the existence of gender-related stereotypes across a range of eight languages but also pioneered a methodology that does not hinge on labor-intensive manual annotations for every language. Finally, through its dual measures, the Skew and Stereotype Methodology unearthed a fascinating interplay between skew and stereotype across models, indicating an inherent trade-off and emphasizing that addressing one might inadvertently exacerbate the other.

6. Conclusion and future works

The recent analysis of gender bias in NLP has illuminated several key areas of concern. Firstly, there is an ethical challenge surrounding the handling of gender, where research often defaults to binary interpretations, neglecting the complexities of non-binary identities. Secondly, there is a disproportionate focus on English and high-resource languages, leading to a narrow view of gender bias and disregarding cultural variances. Thirdly, many NLP models undergo bias testing only post-release, risking unintended societal consequences. The research community must incorporate bias detection during the model creation phase and adopt ethical practices from inception. Lastly, there is a lack of consistency in defining and measuring gender bias, with many studies employing singular, limited definitions. To tackle these issues, the NLP community should prioritize inclusivity, diversify language focuses, embed early-stage bias evaluations, and adopt standardized benchmarks to address gender bias holistically.

This study sheds light on the issue of gender bias in Transformers by undertaking a comprehensive and critical analysis of various works in this field. Through this analysis, we identify the key challenges and limitations in existing research on gender bias in Transformers and opportunities for further investigation. Our study highlights the need for a more nuanced and sophisticated understanding of gender bias in language models and the development of more effective techniques for detecting and mitigating these biases. By critically evaluating existing research, we are able to identify gaps in our knowledge and suggest potential areas for future inquiry. Overall, our linguistic perspective provides a valuable framework for understanding and addressing the problem of gender bias in Transformers. By focusing on the linguistic processes at play, we can develop more targeted and effective interventions to promote gender-fair language models. There are several potential avenues for future research and practical applications in the realm of gender bias in Transformer models. One area of exploration is the development of more sophisticated and nuanced methods for detecting and mitigating gender bias in language models. This could include the creation of new benchmark datasets, as well as the utilization of advanced machine learning techniques and models. Another potential area of application is the deployment of gender-fair language models in real-world settings, such as chatbots or virtual assistants. Such models could play a critical role in promoting inclusivity and diversity in technology by ensuring that language models are free from gender bias and can interact with users in a fair and equitable manner. Overall, the field of gender bias in Transformer models is rapidly evolving, with new research and applications emerging all the time. By continuing to explore these important issues, we can help build a more just and equitable future for all technology users.

7. Ethics and impact statement

Our research focuses on the topic of social biases that are inherently present in huge pre-trained transformer models that are widely accessible and employed. Our findings show that bias is a serious issue that the community has to address and that all pre-trained algorithms currently display some sort of biased gender prediction in otherwise neutral circumstances. Our research also depicts the best-proposed solutions to tackle gender bias in transformers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Anon, 2023. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. (Accessed 3 January 2023).

Asyrofi, M.H., Yang, Z., Yusuf, I.N.B., Kang, H.J., Thung, F., Lo, D., 2022. BiasFinder: Metamorphic test generation to uncover bias for sentiment analysis systems. IEEE Trans. Softw. Eng. 48 (12), 5087–5101. http://dx.doi.org/10.1109/TSE.2021.3136169.

Awasthi, P., Kleindessner, M., Morgenstern, J., 2020. Equalized odds postprocessing under imperfect group information. In: Chiappa, S., Calandra, R. (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol. 108, PMLR, pp. 1770–1780, URL https://proceedings.mlr.press/v108/awasthi20a.html.

Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaparkhi, A., Rosenzweig, J., Sarkar, A., Bangalore, S., 1995. University of pennsylvania: description of the university of pennsylvania system used for MUC-6. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.

Bao, X., Qiao, Q., 2019. Transfer learning from pre-trained BERT for pronoun resolution. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 82–88.

Bartl, M., Nissim, M., Gatt, A., 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. arXiv preprint arXiv:2010.14534.

Basta, C.R.S., Ruiz Costa-Jussà, M., Rodríguez Fonollosa, J.A., 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In: Proceedings of the the Fourth Widening Natural Language Processing Workshop. Association for Computational Linguistics, pp. 99–102.

Beamer, S., Asanović, K., Patterson, D., 2015. The GAP benchmark suite. arXiv preprint arXiv:1508.03619.

Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620.

Bhardwaj, R., Majumder, N., Poria, S., 2021. Investigating gender bias in BERT. Cogn. Comput. 13 (4), 1008–1018.

Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29.

Bordia, S., Bowman, S., 2019. Identifying and reducing gender bias in word-level language models. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 7–15.

Borji, A., 2023. A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494.

Braşoveanu, A.M., Andonie, R., 2020. Visualizing transformers for nlp: A brief survey. In: 2020 24th International Conference Information Visualisation. (IV), IEEE, pp. 270–279.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., Zemel, R., 2019. Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning. PMLR, pp. 803–811.

Budzianowski, P., Vulić, I., 2019. Hello, it's GPT-2-how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In: Proceedings of the 3rd Workshop on Neural Generation and Translation. pp. 15–22.

Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency. In: Proceedings of Machine Learning Research, vol. 81, PMLR, pp. 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Caliskan, A., Ajay, P.P., Charlesworth, T., Wolfe, R., Banaji, M.R., 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 156–170.

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

de Vassimon Manela, D., Errington, D., Fisher, T., van Breugel, B., Minervini, P., 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In: EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. Association for Computational Linguistics, pp. 2232–2242.

Floridi, L., Chiriatti, M., 2020. GPT-3: Its nature, scope, limits, and consequences. Minds Mach. 30, 681–694.

Garg, P., Villasenor, J., Foggo, V., 2020. Fairness metrics: A comparative analysis. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, pp. 3662–3666.

Garimella, A., Banea, C., Hovy, D., Mihalcea, R., 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 3493–3498. http://dx.doi.org/10.18653/v1/P19-1339, URL https://aclanthology.org/P19-1339.

Gillioz, A., Casas, J., Mugellini, E., Abou Khaled, O., 2020. Overview of the transformer-based models for NLP tasks. In: 2020 15th Conference on Computer Science and Information Systems. (FedCSIS), IEEE, pp. 179–183.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al., 2021. Pre-trained models: Past, present and future. AI Open 2, 225–250.

He, P., Liu, X., Gao, J., Chen, W., 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A., 2018. Women also snowboard: Overcoming bias in captioning models. In: Proceedings of the European Conference on Computer Vision. (ECCV), pp. 771–787.

Hirota, Y., Nakashima, Y., Garcia, N., 2022. Quantifying societal bias amplification in image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13450–13459.

Hovy, D., Bianchi, F., Fornaciari, T., 2020. "You sound just like your father" commercial machine translation systems include stylistic biases. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1686–1690.

Hovy, D., Prabhumoye, S., 2021. Five sources of bias in natural language processing. Lang. Linguist. Compass 15 (8), e12432.

Jangir, R., Hansen, N., Ghosal, S., Jain, M., Wang, X., 2022. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. IEEE Robot. Autom. Lett. 7 (2), 3046–3053.

Joniak, P., Aizawa, A., 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. arXiv preprint arXiv:2207.02463.

Kaneko, M., Imankulova, A., Bollegala, D., Okazaki, N., 2022. Gender bias in masked language models for multiple languages. arXiv preprint arXiv:2205.00551.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. ACM Comput. Surv. (CSUR) 54 (10s), 1–41.

Kiritchenko, S., Mohammad, S., 2018a. Examining gender and race bias in two hundred sentiment analysis systems. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 43–53.

Kiritchenko, S., Mohammad, S., 2018b. Examining gender and race bias in two hundred sentiment analysis systems. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, New Orleans, Louisiana, pp. 43–53. http://dx.doi.org/10.18653/v1/S18-2005, URL https://aclanthology.org/S18-2005.

Li, B., Peng, H., Sainju, R., Yang, J., Yang, L., Liang, Y., Jiang, W., Wang, B., Liu, H., Ding, C., 2021. Detecting gender bias in transformer-based models: A case study on bert. arXiv preprint arXiv:2110.15733.

Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., Solti, I., 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J. Am. Med. Inform. Assoc. 21 (3), 406–413.

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

Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A., 2020. Gender bias in neural natural language processing. In: Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday. Springer, pp. 189–202.

Lund, B.D., Wang, T., 2023. Chatting about chatgpt: how may AI and GPT impact academia and libraries? Library Hi Tech News.

Luo, W., 2019. Encoder-Decoder Based Neural Machine Translation (Ph.D. thesis).
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54 (6), 1–35.
Nadeem, A., Abedin, B., Marjanovic, O., 2020. Gender bias in AI: A review of contributing factors and mitigating strategies.
Nadeem, A., Marjanovic, O., Abedin, B., et al., 2022. Gender bias in AI-based decision-making systems: A systematic literature review. Australas. J. Inf. Syst. 26.
Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J.C., Alonso, A., 2023. Linguistic ambiguity analysis in ChatGPT. arXiv preprint arXiv:2302.06426.
Ozieblowska, B., 1994. Generic Pronouns in Current Academic Writing. University of Glasgow (United Kingdom).
Papakyriakopoulos, O., Hegelich, S., Serrano, J.C.M., Marco, F., 2020. Bias in word embeddings. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 446–457.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.
Park, J.H., Shin, J., Fung, P., 2018. Reducing gender bias in abusive language detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2799–2804.
Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., Kamar, E., 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7. pp. 125–134.
Prates, M.O., Avelar, P.H., Lamb, L.C., 2020. Assessing gender bias in machine translation: A case study with google translate. Neural Comput. Appl. 32, 6363–6381.
Reimers, N., Gurevych, I., 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), pp. 3982–3992.
Robinson, R., 2021. Assessing gender bias in medical and scientific masked language models with StereoSet. arXiv preprint arXiv:2111.08088.
Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M., 2021. Gender Bias in Machine Translation. Trans. Assoc. Comput. Linguist. 9, 845–874. http://dx.doi.org/10.1162/tacl_a_00401.
Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P., et al., 2022. Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publ. 1270, 1–77.
Silva, A., Tambwekar, P., Gombolay, M., 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2383–2389.
Simundic, A.-M., 2013. Bias in research. Biochem. Med. 23 (1), 12–15.
Stanovsky, G., Smith, N.A., Zettlemoyer, L., 2019. Evaluating gender bias in machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1679–1684.
Sweeney, L., 2013. Discrimination in online ad delivery. Commun. ACM 56 (5), 44–54.
Tan, Y.C., Celis, L.E., 2019. Assessing social and intersectional biases in contextualized word representations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf.
Tang, R., Du, M., Li, Y., Hu, X., 2020. Mitigating gender bias in captioning systems.
Thelwall, M., 2018. Gender bias in sentiment analysis. Online Inf. Rev..
Turney, P., 1995. Bias and the quantification of stability. Mach. Learn. 20, 23–33.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.
Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S., 2020. Investigating gender bias in language models using causal mediation analysis. Adv. Neural Inf. Process. Syst. 33, 12388–12401.
Wang, Y., Redmiles, D., 2019. Implicit gender biases in professional software development: An empirical study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society. (ICSE-SEIS), IEEE, pp. 1–10.
Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics, San Diego, California, pp. 88–93. http://dx.doi.org/10.18653/v1/N16-2013, URL https://aclanthology.org/N16-2013.
West, M., Kraut, R., Chew, E., H., 2019. I'D Blush if I Could: Closing Gender Divides in Digital Skills Through Education. UNESCO.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
Wolfe, R., Caliskan, A., 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 518–532.
Wołk, K., Marasek, K., 2015. Enhanced bilingual evaluation understudy. arXiv preprint arXiv:1509.09088.
Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.-W., 2019. Gender bias in contextualized word embeddings. In: Proceedings of NAACL-HLT. pp. 629–634.
Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.-W., 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 15–20. http://dx.doi.org/10.18653/v1/N18-2003, URL https://aclanthology.org/N18-2003.