# Evaluating Short-Term Temporal Fluctuations of Social Biases in Social Media Data and Masked Language Models

Anonymous ACL submission

## Abstract

Social biases such as gender or racial biases have been reported in language models (LMs), including Masked Language Models (MLMs). Given that MLMs are continuously trained with increasing amounts of additional data collected over time, an important yet unanswered question is how the social biases encoded with MLMs vary over time. In particular, the number of social media users continues to grow at an exponential rate, and it is a valid concern for the MLMs trained specifically on social media data whether their social biases (if any) would also amplify over time. To empirically analyse this problem, we use a series of MLMs pretrained on chronologically ordered temporal snapshots of corpora. Our analysis reveals that, although social biases are present in all MLMs, most types of social bias remain relatively stable over time (with a few exceptions). To further understand the mechanisms that influence social biases in MLMs, we analyse the temporal corpora used to train the MLMs. Our findings show that some demographic groups, such as *male*, obtain higher preference over the other, such as *female* on the training corpora constantly.[1]

## 1 Introduction

Despite their usage in numerous NLP applications, MLMs such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) tend to encode discriminatory social biases expressed in human-written texts in the training corpora (Kurita et al., 2019; Zhou et al., 2022; Kaneko et al., 2022). For example, if a model is given "*[MASK] is a nurse.*" as the input, a gender biased MLM would predict "*She*" with a higher likelihood score than for "*He*" when filling the [MASK]. Such social biases can result in unfavourable experiences for some demographic groups in certain applications. Continuous use of biased models has the potential to amplify biases and unfairly discriminate against users belonging to particular demographic groups. MLMs are increasingly used in real-world applications such as text generation (Liang et al., 2023), recommendation systems (Malkiel et al., 2020; Kuo and Li, 2023), search engines (Achsas et al., 2022; Li et al., 2023) and dialogue systems (Song et al., 2021; Park et al., 2022). Therefore, it is crucial to study how MLMs potentially shape social biases.

On the other hand, social biases may change due to societal changes, cultural shifts and technological advancements. MLMs have been trained on ever-increasing massive corpora, often collected from the Web. In particular, posts on social media, such as but not limited to Reddit and X (former Twitter), have been used to train MLMs. Social biases contained in the training data are inadvertently learned and perpetuated by MLMs. At the time of writing, there are 5.07 billion social media users worldwide with 259 million new users joining since this time in 2023.[2] Given this rapid increase and the significance of social media data as a source for training MLMs, an open question is **whether LMs trained on social media data continue to demonstrate increasing levels of social biases.**

To answer this question, we investigate multiple MLMs pretrained on snapshots of corpora collected from X at different points in time and evaluate the social biases in those MLMs using multiple benchmark datasets. We evaluate different types of social biases and observe that the overall bias tends to be stable over time, however, certain types of biases, such as race, skin color, religion, and sexual orientation, exhibit fluctuation over time. Based on the experimental results, we note that relying exclusively on the overall bias score can be misleading when evaluating social bias in MLMs, which highlights the importance of evaluating individual bias

[1] An anonymised version of the code is submitted to ARR and will be publicly released upon paper acceptance. Note that we are mainly using existing evaluation code in this submission (which is referred to in the paper), and thus we do not present a new model or evaluation metric.

[2] https://datareportal.com/social-media-users

scores before deploying a model in downstream applications. Note that we primarily investigate whether language models (LMs) trained on social media data exhibit increasing levels of social biases over time in this paper. Our focus is on examining the trends in temporal variations of social biases in both models and datasets. Exploring the underlying causes could lead to sociologically oriented experiments and research questions, which are beyond the scope of this NLP-focused study.

## 2 Related Work

**Social Biases in NLP.** Social biases in NLP were first drawn to attention by Bolukbasi et al. (2016), with the famous analogy "*man is to computer programmer as woman is to homemaker*" provided by static word embeddings. To evaluate social biases in word embeddings, word Embedding Association Test (**WEAT**; Caliskan et al., 2017a) was introduced to measure the bias between two sets of target terms with respect to two sets of attribute terms. Subsequently, Word Association Test (**WAT**; Du et al., 2019) was proposed to compute a gender information vector for each word within an association graph (Deyne et al., 2019) through the propagation of information associated with masculine and feminine words. Follow-up studies investigate social biases in additional models (Liang et al., 2020a,b; Zhou et al., 2022) and languages (McCurdy and Serbetci, 2020; Lauscher et al., 2020; Reusens et al., 2023; Zhou et al., 2023).

In contrast, alternative research focuses on social biases in various downstream applications. Kiritchenko and Mohammad (2018) assessed gender and racial biases across 219 automatic sentiment analysis systems, revealing statistically significant biases in several of these systems. Díaz et al. (2018) investigated age-related biases in sentiment classification and found that many sentiment analysis systems, as well as word embeddings, encode significant age bias in their outputs. Savoldi et al. (2021) studied gender biases and sentiment biases associated with person name translations in neural machine translation systems.

Current bias evaluation methods use different approaches, including pseudo-likelihood. (Kaneko and Bollegala, 2022), cosine similarity (Caliskan et al., 2017b; May et al., 2019), inner-product (Ethayarajh et al., 2019), among others. Independently of any downstream tasks, intrinsic bias evaluation measures (Nangia et al., 2020; Nadeem et al., 2021; Kaneko and Bollegala, 2022) assess social biases in MLMs on a standalone basis. Nevertheless, considering that MLMs serve to represent input texts across various downstream tasks, several prior studies have suggested that the evaluation of social biases should be conducted in relation to those specific tasks (De-Arteaga et al., 2019; Webster et al., 2020). Kaneko and Bollegala (2021) demonstrated that there is only a weak correlation between intrinsic and extrinsic social bias evaluation measures. In this paper, we use AULA which is an intrinsic measure for evaluating social biases in MLMs.

Various debiasing methods have been proposed to mitigate social biases in MLMs. Zhao et al. (2019) proposed a debiasing method by swapping the gender of female and male words in the training data. Webster et al. (2020) showed that dropout regularisation can reduce overfitting to gender information, thereby can be used for debiasing pretrained language models. Kaneko and Bollegala (2021) proposed a method for debiasing by orthogonalising the vectors representing gender information with the hidden layer of a language model given a sentence containing a stereotypical word. Our focus in this paper is the evaluation of social biases rather than proposing bias mitigation methods.

**Temporal Variations in MLMs.** Diachronic Language Models that capture the meanings of words at a specific timestamp have been trained using historical corpora (Qiu and Xu, 2022; Loureiro et al., 2022a). Rosin and Radinsky (2022) introduced a temporal attention mechanism by extending the self-attention mechanism in transformers. They took into account the time stamps of the documents when calculating the attention scores. Tang et al. (2023b) proposed an unsupervised method to learn dynamic contextualised word embeddings via time-adapting a pretrained MLM using prompts from manual and automatic templates. Aida and Bollegala (2023) proposed a method to predict the semantic change of words by comparing the distributions of contextualised embeddings for the word between two corpora sampled at different timestamps. Tang et al. (2023a) used word sense distributions to predict semantic changes of words in English, German, Swedish and Latin.

On the other hand, Zeng et al. (2017) learned *socialised* word embeddings by taking into account both the personal characteristics of language used by a social media user and the social relationships of that user. Welch et al. (2020) learned demographic word embeddings, covering attributes such as age, gender, location and religion. Hofmann

et al. (2021) demonstrated that temporal factors exert a more significant influence than socio-cultural factors in determining the semantic variations of words. However, to the best of our knowledge, the temporal changes of social biases in MLMs remains understudied, and our focus in this paper is to fill this gap.

## 3 Temporal Data and Models

To investigate the temporal variant of social biases appearing in the corpora, we retrieve the posts on X with different timestamps. Furthermore, we take into account the MLMs trained on those temporal corpora to study how MLMs potentially shape social biases from these corpora. In this section, we describe the temporal data and the MLMs that we used in the paper.

### 3.1 Temporal Corpora

We use the snapshots of corpora from X across a two-year time span – from the year 2020 to 2022, collected using Twitter's Academic API.[3] To obtain a sample that is reflective of the general conversation of people's daily lives on social media, we follow the collection process from Loureiro et al. (2022b) in order to collect a diverse corpus while avoiding duplicates and spam.

Specifically, we use the API to retrieve tweets using the most frequently used stopwords,[4] capturing a predetermined number of tweets at intervals of 5 minutes. This process is carried out for each hour and every day, spanning a specific quarterly period in the year. In addition, we leverage specific flags supported by the API to exclusively fetch tweets in English, disregarding retweets, quotes, links, media posts, and advertisements. Assuming bots are among the most active users, we eliminate tweets from the top $1\%$ of the most frequent posters.

To ensure the dataset remains free of duplicates, we eliminate both exact and near-duplicate tweets. Specifically, we first convert tweets to lowercase and remove punctuation. Then we identify near-duplicates by generating hashes using MinHash (Broder, 1997) with 16 permutations. Finally, non-verified user mentions are substituted by a generic placeholder (@user). The statistics of temporal corpora can be found in Appendix A.

[3]Twitter Academic API was interrupted in 2023, and that is the reason why our data collection was interrupted after the end of 2022.

[4]We select the top 10 ones from https://raw.githubusercontent.com/first20hours/google-10000-english/master/google-10000-english.txt

### 3.2 Models trained on Different Timestamps

To investigate whether social biases in MLMs exhibit temporal variation, we evaluate social biases in MLMs that are trained on corpora sampled from different timestamps. Specifically, we select the pre-trained TimeLMs[5] (Loureiro et al., 2022b), which are a set of language models trained on diachronic data from X. TimeLMs are continuously trained using data collected from X, starting with the initial RoBERTa base model (Liu et al., 2019). The base model of TimeLMs is first trained with data until the end of 2019. Since then, subsequent models have been routinely trained every three months, building upon the base model. To ensure the models trained on the corpora sampled with different timestamps are with the same setting (i.e., with incremental updates), we discard the base model trained until 2019 and select the models trained with the temporal corpora described in § 3.1.

To investigate the fluctuations in social biases in MLMs over time, we require a series of pre-trained MLMs of the same architecture, trained on corpora sampled at different timestamps. To the best of our knowledge, such MLMs based on architectures other than RoBERTa do not currently exist. Furthermore, training these temporal models from scratch, such as pre-training MLMs with a different architecture, is computationally expensive and time-consuming. For instance, training a RoBERTa base temporal model takes approximately 15 days on 8 NVIDIA V100 GPUs. Given that pretrained temporal MLMs based on models other than RoBERTa are not available, and Zhou et al. (2023) show that various underlying factors differentially impact social biases in MLMs, our approach focuses on using models that have been continuously trained from an existing RoBERTa base checkpoint. This strategy maintains consistency in model settings, which aids in accurately assessing how MLMs reflect the temporal variations in social biases.

## 4 Experimental Setting

Our goal in this paper is to study whether MLMs capture temporal changes in social biases, following the same patterns observed in the biases present in training corpora. For this purpose, we evaluate social biases in MLMs and compare the biases observed in training corpora.

[5]https://github.com/cardiffnlp/timelms

### 4.1 Bias Evaluation Metrics

To investigate the social biases within MLMs, we compute social bias scores of TimeLMs using All Unmasked Likelihood with Attention weights (AULA; Kaneko and Bollegala, 2022). This metric evaluates social biases by using MLM attention weights to reflect token significance. AULA has proven to be more robust against frequency biases in words for evaluating social biases in MLMs and offers more reliable evaluations in comparison to alternative metrics when assessing social biases in MLMs (Kaneko et al., 2023). Further details on the computation of AULA are shown in Appendix B

### 4.2 Benchmarks

We perform experiments on the two most commonly used benchmark datasets used to evaluate social biases in MLMs.

**CrowS-Pairs** (Nangia et al., 2020). proposed Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs), which is designed to explore stereotypes linked to historically disadvantaged groups. It is a crowdsourced dataset annotated by workers in the United States and contains nine social bias categories: race, gender, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status/occupation. In the CrowS-Pairs dataset, test instances comprise pairs of sentences, where one sentence is stereotypical and the other is anti-stereotypical. Annotators are instructed to generate examples that indicate stereotypes by contrasting historically disadvantaged groups with advantaged groups.

**StereoSet** (Nadeem et al., 2021). created StereoSet, which includes associative contexts encompassing four social bias types: race, gender, religion, and profession. StereoSet incorporates test instances at both intrasentence and intersentence discourse levels. They introduced a Context Association Test (CAT) to assess both the language modelling ability and the stereotypical biases of pretrained MLMs. Specifically, when presented with a context associated with a demographic group (e.g., female) and a bias type (e.g., gender), three distinct labels are provided to instantiate its context, corresponding to a stereotypical, anti-stereotypical, or unrelated association.

We use the social bias evaluation tool released by Kaneko and Bollegala (2022)[6] with its default settings for all evaluations reported in this paper.

[6] https://github.com/kanekomasahiro/evaluate_bias_in_mlm

## 5 Temporal Variation of Social Biases

In this section, we describe the key findings of our paper, presenting a comprehensive analysis and interpretation of the results.

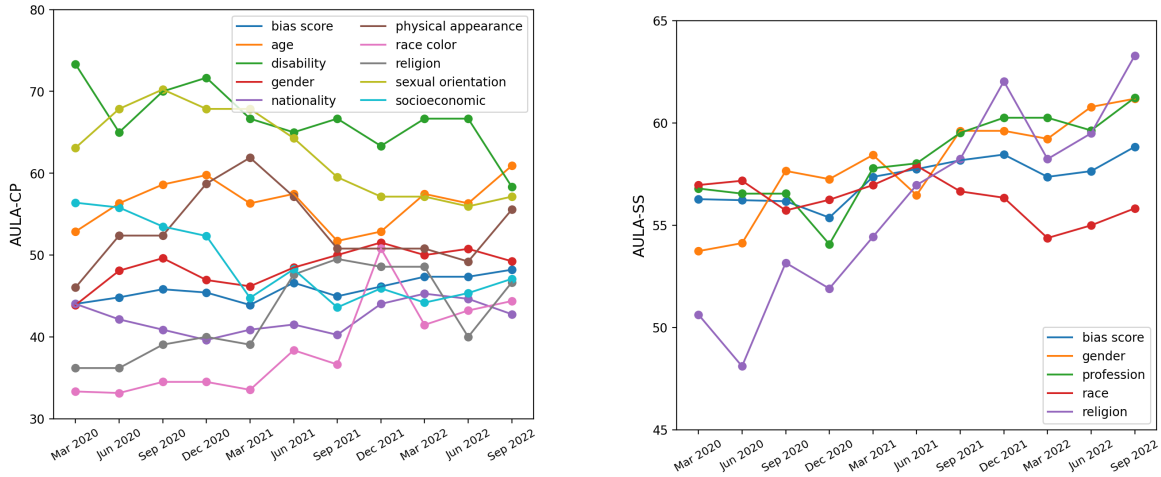
### 5.1 Biases in MLMs

Figure 1 shows the changes of bias scores for different bias types in TimeLMs over the period from March 2020 to September 2022 computed by AULA on both CrowS-Pairs and StereoSet datasets. It is noticeable that different types of biases within TimeLMs change over time. The overall bias scores exhibit minimal changes over time compared to other types of biases in both datasets. This result suggests that even when there is no overall social bias reported by a metric, an MLM can still be biased with respect to a subset of the bias types. Therefore, it is important to carefully evaluate bias scores per each bias type before an MLM is deployed in downstream applications.

| | mean | lower/upper | SE | SD |
|---|---|---|---|---|
| **CrowS-Pairs** | | | | |
| OVERALL BIAS | 45.88 | 45.21/46.55 | 0.41 | 1.41 |
| race-color | 38.53 | 36.19/41.88 | 1.68 | 5.77 |
| sexual-orientation | 62.55 | 60.06/65.15 | 1.54 | 5.36 |
| religion | 42.86 | 40.35/45.45 | 1.52 | 5.30 |
| socioeconomic | 48.84 | 46.78/51.32 | 1.37 | 4.79 |
| appearance | 53.25 | 51.23/55.70 | 1.33 | 4.62 |
| disability | 66.67 | 64.70/68.49 | 1.17 | 4.08 |
| age | 56.42 | 54.86/57.68 | 0.85 | 2.93 |
| gender | 48.61 | 47.40/49.55 | 0.64 | 2.23 |
| nationality | 42.37 | 41.51/43.28 | 0.55 | 1.91 |
| **StereoSet** | | | | |
| OVERALL BIAS | 57.23 | 56.70/57.74 | 0.31 | 1.09 |
| religion | 56.04 | 53.62/58.34 | 1.39 | 4.81 |
| gender | 58.00 | 56.72/59.07 | 0.71 | 2.47 |
| profession | 58.24 | 57.15/59.22 | 0.62 | 2.15 |
| race | 56.28 | 55.77/56.73 | 0.29 | 1.02 |

Table 1: Confidence intervals and standard errors are computed using bootstrapping test for each bias type on the CrowS-Pairs and StereoSet benchmarks. SE and SD represent standard error and standard deviation, respectively. Lower/upper indicates the lower/upper bound of the confidence intervals. In each dataset, different bias types are sorted in the descending order of their SD.

When evaluating on CrowS-Pairs, we observe that both disability and sexual orientation biases consistently receive bias scores above 50. This indicates a consistent inclination of these two biases toward stereotypical examples over a span of two




(a) CrowS-Pairs (b) StereoSet

Figure 1: Social bias scores across time for different types of biases computed using the AULA metric. Results evaluated on the CrowS-Pairs and StereoSet datasets are shown respectively on the left and right. The 'bias score' (in dark blue) indicates the overall bias score.

years. Conversely, religion and nationality exhibit a consistent inclination toward anti-stereotypical examples over time. In terms of the evaluation on StereoSet, most types of biases exhibit stereotypical tendencies, except the religious bias in June 2020, which leaned toward anti-stereotypical examples. In particular, the religious biases have increased from 51 to 63 over the two year period from 2020 to 2022. This finding highlights the nuanced nature of different types of biases and their variations across different contexts, encouraging future research aimed at establishing a benchmark that equally considers different types of biases (Blodgett et al., 2021). However, our primary focus is on investigating the temporal fluctuations of social biases in MLMs, and as such, the specific direction of different biases presenting differently on the evaluation datasets is out of the scope of this paper.

**Statistical indicators of bias fluctuation changes.** To further validate the consistency of the aforementioned observations, we use the bootstrapping significance test (Tibshirani and Efron, 1993) to the temporal variation of different social bias types. Specifically, given a bias type, we first compute the AULA score over the entire dataset at a particular time point, resulting in a series of data points, each one corresponding to a particular time point, and we report the average and standard deviation of that score along with its confidence interval and standard error computed using bootstrapping. Bootstrapping is a statistical technique which uses random sampling with replacement. By measuring the properties when sampling from an approximating distribution, bootstrapping estimates the properties of an estimand (e.g., variance). We implement bootstrapping using the SciPy[7] at 0.9 confidence level to compute the confidence intervals, while setting other parameters to their defaults.

Table 1 shows the result. In CrowS-Pairs, the bias types such as sexual orientation, physical appearance, disability, and age manifest biases mostly toward stereotypical examples (i.e., the mean of their bias scores are above 50), while biases associated with race colour, religion, socioeconomic, gender and nationality tend to have biases toward anti-stereotypical examples (i.e., the mean of their bias scores are below 50). On the other hand, race colour reports the highest standard error, indicating that it is the most fluctuating bias type over time.

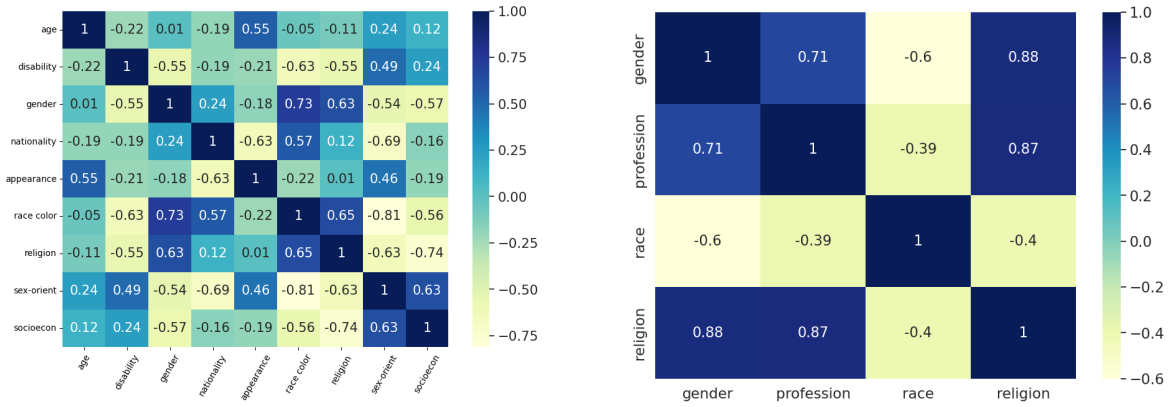
In StereoSet, we observe all the types of biases exhibit biases toward stereotypical examples. Moreover, religion is the most fluctuating bias over time compared to other types of biases, while racial bias does not change much over time. Note that the CrowS-Pairs dataset assesses race colour bias, specifically concentrating on the skin colour associated with race, which is different from the race bias considered in StereoSet.

## 5.2 Correlations between Bias Types

To investigate whether the change in one type of bias influences other types, we compute the Pearson correlation coefficient ($r$) for each pair of bias types. We use the SciPy library[8] with the default

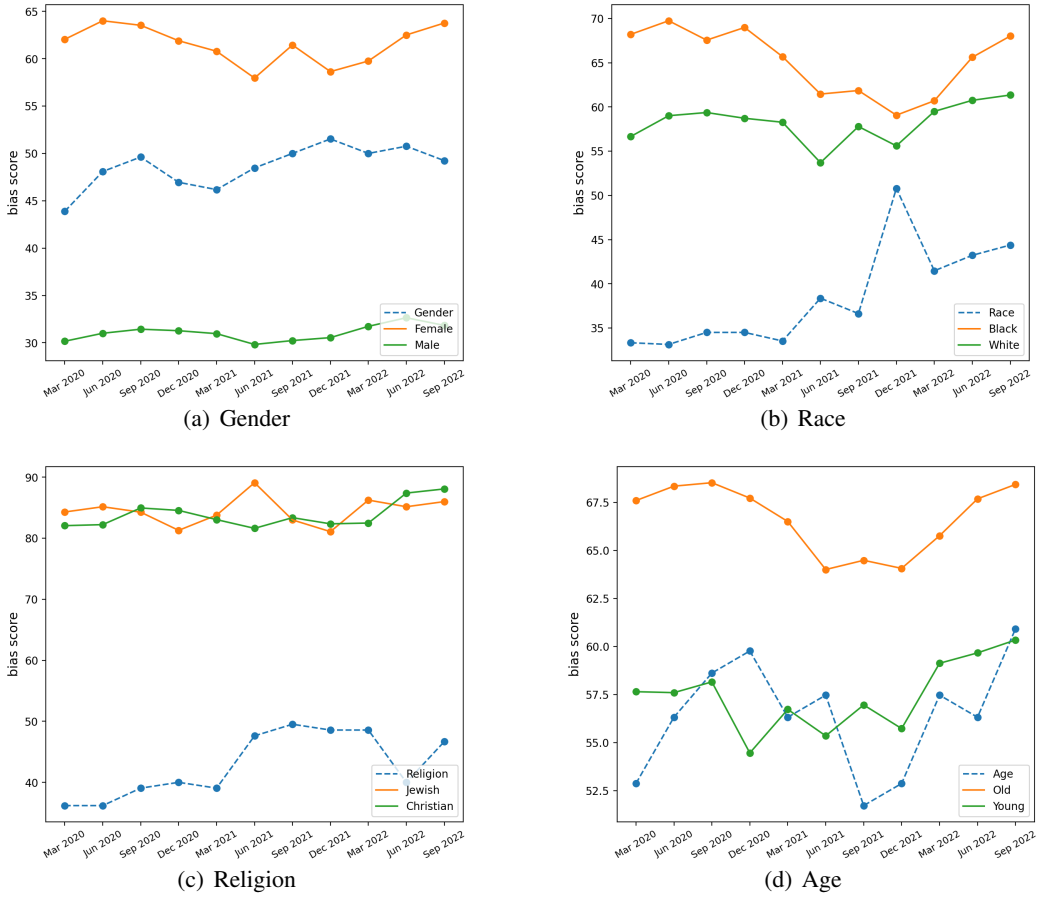
[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html




(a) CrowS-Pairs (b) StereoSet

Figure 2: Pearson correlation coefficient of each pair of bias types. Results on the CrowS-Pairs and StereoSet datasets are shown respectively on the left and right.




(a) Gender (b) Race

(c) Religion (d) Age

Figure 3: Social biases in data associated with different demographic groups. A sentiment classifier is used to determine whether a tweet associated with a particular demographic group conveys positive or negative sentiment. Dash line represents the bias scores computed using (2) on CrowS-Pairs, while solid lines show bias scores computed using (1), respectively.

setting for doing so and show the results in Figure 2. When evaluating on CrowS-Pairs, race color and gender biases have the highest correlation (i.e., 0.73) compared to other bias pairs, whereas race color obtains the lowest correlation (i.e., -0.81) with sexual orientation. Moreover, strong positive correlations (i.e., $r > 0.65$) exist among pairs such as race colour vs. gender and race colour vs, religion, while sexual orientation vs. race colour, sexual orientation vs. nationality and socioeconomic

vs. religion obtain strong negative correlation (i.e., $r < -0.65$).

As far as StereoSet is concerned, we observe that the pairs such as profession vs. gender, religion vs, gender, and religion vs, profession exhibit strong positive correlations (i.e., $r > 0.65$), while race vs. gender, race vs. profession, as well as religion vs. race, manifest negative correlations.

### 5.3 Biases in Data

To study the presence of biases related to a certain demographic group in the training corpus and the extent to which an MLM learns these biases during pre-training, we measure different types of social biases appearing in the corpus. Following prior work that evaluates bias in words using their association to pleasant vs. unpleasant words (Caliskan et al., 2017a; Du et al., 2019), we evaluate the bias score of a demographic group $\mathcal{D}$ by considering its members $x \in \mathcal{D}$, and their association with positive and negative contexts.

However, instead of relying on a fixed set of pleasant/unpleasant words, which is both limited and the occurrence of a single word could be ambiguous, we use sentiment classification as a proxy for eliciting such pleasant (expressed by a positive sentiment) and unpleasant (expressed by a negative sentiment) judgements. For this purpose we use the sentiment classification model fine-tuned on TweetEval (Barbieri et al., 2020),[9] which associates each tweet with a positive, negative or neutral sentiment. According to Kiritchenko and Mohammad (2018), some sentiment analysis models show biases, particularly related to race more than gender. In this paper, we specifically focus on evaluating biases using a state-of-the-art sentiment analysis model, according to the TweetEval benchmark, that has been fine-tuned on tweets to minimise biases that could arise from varied datasets. It is important to note that our analysis does not extend to comparing biases across different sentiment analysis models, which is beyond the scope of this paper.

Given a word $x \in \mathcal{D}$ that occurs in a sentence $S$, we use the *negativity score* to measure the social biases in the training data. The negativity score of the group $\mathcal{D}$ is defined by (1).

$$\text{Score} = 100 \times \frac{\sum_{x \in \mathcal{D}} S_{\text{n}}(x)}{\sum_{x \in \mathcal{D}} S_{\text{p}}(x) + S_{\text{n}}(x)} \quad (1)$$

Here, $S_{\text{p}}(x)$ and $S_{\text{n}}(x)$ represent the number of times that $S$ is classified as respectively *positive* or *negative* by a sentiment classifier given the word $x$ appear in the sentence $S$. Similar to the bias score computed using AULA, an unbiased dataset will return a bias score of 50, while greater and lower than 50 indicates the bias toward stereotypical and anti-stereotypical examples, respectively.

We select four types of biases and categorise them according to the magnitude of changes over time. Based on the results shown in Table 1, we focus on those with minimal changes (i.e., standard error less than 1.00), which are age and gender biases, and those with more pronounced changes (i.e., standard error greater than 1.00), which are race colour and religion for evaluation. Note that the racial and religious biases in CrowS-Pairs and StereoSet are sub-categorised and cover more than two demographic groups. However, in the following evaluation, we take into account two demographic groups for each of the bias types.

**Gender Bias.** We retrieve the top-50 *male* and *female* names respectively from Name Census: United States Demographic,[10] which contains the most popular baby names from 1880 to the latest available data in 2022. These names are directly sourced from Social Security card applications submitted for the births in the United States. The detailed list of the words we used for the demographic descriptor words for gender bias can be found in § C.1.

Figure 3(a) shows the results. The *male* category consistently obtains a low negativity score (i.e., $< 35$), while *female* returns high negativity scores (i.e., $> 55$) across time. This indicates that the words in the *male* group constantly exhibit a strong association with positive tweets compared to the *female* group. Moreover, the *male* bias exhibits stability over time, whereas *female* bias shows more fluctuations.

**Racial Bias.** To evaluate racial bias occurring in training corpora, we select the names that are associated with being African American and European American from the work by Kiritchenko and Mohammad (2018), consisting of 20 names in each of the demographic groups. The lists of words representing *White* and *Black* races used in our paper are shown in § C.2.

From Figure 3(b) we observe that both *Black* and *White* biases reduce from June 2020 to June 2021, while both increase from December 2021 to September 2022. Conversely, the overall racial

[9] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[10] https://namecensus.com/baby-names/

bias contains a different trend. The overall racial bias remains stable until March 2021. In addition, both *Black* and *White* biases have higher levels of social biases toward stereotypical examples, while the overall racial bias tends to be anti-stereotypical, except in December 2021, when it reaches its peak.

**Religious Bias.** In terms of religious bias, we consider the terms associated with *Jewish* and *Christian* identities and choose terms listed as the demographic identity labels from AdvPromptSet (Esiobu et al., 2023), and the phrases related to demographic groups are listed in § C.3.

The result of the religious bias scores as well as the negativity scores associated with *Christian* and *Jewish* identities are shown in Figure 3(c). Regarding biases associated with *Jewish* and *Christian* in the data, we observe that both biases obtain high levels of social bias toward stereotypes. However, the general religious bias in MLMs demonstrates a lower degree of social biases, primarily towards anti-stereotypes over time. On the other hand, the *Christian* bias is more stable compared to *Jewish* and overall religious biases.

**Age Bias.** For the age bias, we consider the demographic categories of *young* and *old*. Therefore, we use the descriptor terms in HOLISTICBIAS Smith et al. (2022), and the list of the terms associated with young and old can be found in § C.4.

Figure 3(d) shows the bias associated with *young* and *old* demographic groups along with the overall age bias over time. We observe that from December 2021 to March 2022, the negativity score associated with the *old* group increases along with the overall age bias. However, we can observed a marked difference in terms of absolute values, with the negativity score for the *old* group being generally much larger.

**Control Analysis.** To further verify whether social biases also vary independently of time, we conduct a control analysis by randomly sampling a subset of a corpus within the same time period. Specifically, we consider social biases associated with *female* and *male* and randomly sample 1/5 of the tweets from January to March 2020 for 5 times and compute the standard deviation of *female* and *male* bias scores over these samples.

The standard deviations of both *female* and *male* biases in a corpus sampled with the same timestamp are 0.16 and 0.19, respectively, which are much lower than the standard deviations of *female* (i.e., 2.03) and *male* biases (i.e., 0.84) across time. This indicates that the temporal aspect has a more pronounced effect on social biases, showing that social biases do not vary independently of time. The details of the results for social biases in random sample subsets and in the temporal corpora are shown in Appendix D.

### 5.4 Comparison with temporal bias fluctuations in historical data

To further investigate the fluctuations of social biases present in corpora with a longer time span, we apply the same setting as in § 4 on COHABERT,[11] which is a series of RoBERTa base models that are continuously trained on COHA (Davies, 2015). COHA is the largest structured corpus of historical English. The COHABERT models have been trained over a long period, spanning from the year 1810 to 2000.

Due to space limitations, the results for different bias types and their historical fluctuations are shown in the appendix (§ E.1 and § E.2, respectively). Overall, biases show more fluctuations over a longer time span (i.e., exhibiting higher standard deviations over time) than over a shorter one. Comparing the different bias types within COHABERT models, we observe a similar trends over time, demonstrating that overall bias scores remain relatively stable compared to specific bias types across both CrowS-Pairs and StereoSet. Specifically, the overall bias in COHA produced standard deviations of 1.11 in StereoSet and 3.59 in CrowS-Pairs when measured in 10-year span periods. Sexual orientation is the most fluctuating bias type in CrowS-Pairs, whereas religion shows the most variability over time in StereoSet.

## 6 Conclusion

We studied the temporal variation of social biases appearing in the data as well as in MLMs. We conducted a comprehensive study using various pretrained MLMs trained on different snapshots of datasets collected at different points in time. While social biases associated with some demographic groups undergo changes over time, the results show that the overall social biases, as captured by language models and as analysed on the underlying corpora, remain relatively stable. Therefore, using the overall bias score without considering different bias types to indicate social biases present in MLMs can be misleading. We encourage future research to consider different types of biases for study, where these biases can be more pronounced.

[11] https://github.com/seongmin-mun/COHABERT

## 7 Limitations

This paper studies the temporal variation of social biases in datasets as well as in MLMs. In this section, we highlight some of the important limitations of this work. We hope this will be useful when extending our work in the future by addressing these limitations.

As described in § 3.2, our main results are based on the RoBERTa base models trained with temporal corpora. This is limited by the availability of language models trained on different time periods. Related to this, the evaluation in this paper is limited to the English language and we only collect temporal corpora on X. Extending the work to take into account models with different architectures for comparison and the study to include multiple languages as well as collecting data from different social media platforms will be a natural line of future work.

As mentioned in § 5.3, certain sentiment analysis models exhibit biases. These biases in such models are more commonly found in relation to race compared to gender. In this paper, we measure biases in data by only taking into account one RoBERTa based sentiment analysis model trained on tweets. However, comparing biases in different sentiment analysis models is out of the scope of this paper.

In this paper, we narrow down our focus to evaluate the intrinsic social biases captured by MLMs. However, there are various extrinsic bias evaluation datasets existing such as BiasBios (De-Arteaga et al., 2019), STS-bias (Webster et al., 2020), NLI-bias (Dev et al., 2020). A logical next step for our research would be to extend our work and assess the extrinsic biases in MLMs.

Due to the computational costs involved when training MLMs, we conduct a control experiment to investigate whether social biases vary independently of time with the focus on biases in data. However, it remains to be evaluated whether the similar trend can be observed for the biases in MLMs.

## 8 Ethical Considerations

In this paper, we aim to investigate whether social biases in datasets and MLMs exhibit temporal variation. Although we used datasets collected from X, we did not annotate nor release new datasets as part of this research. Specifically, we refrained from annotating any datasets ourselves in this study. Instead, we utilised corpora and benchmark datasets that were previously collected, annotated, and consistently employed for evaluations in prior research. To the best of our knowledge, no ethical issues have been reported concerning these datasets. All the data utilised from X has been anonimized, excluding all personal information and only retaining the text in the post, where user mentions were also removed.

The gender biases considered in the bias evaluation datasets in this paper only consider binary gender. However, non-binary genders are severely lacking representation in the textual data used for training MLMs (Dev et al., 2021). Moreover, non-binary genders are frequently associated with derogatory adjectives. It is crucial to evaluate social bias by considering non-binary gender.

## References


Sanae Achsas et al. 2022. Academic aggregated search approach based on bert language model. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, pages 1–9.

Taichi Aida and Danushka Bollegala. 2023. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In *Proc. of the Findings of 61st Annual Meeting of the Association for Computational Linguistics*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, pages 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81.

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

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings.*

*Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. pages 21–29. https://doi.org/10.1109/SEQUEN.1997.666900.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186. https://www.science.org/doi/abs/10.1126/science.aal4230.

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

Mark Davies. 2015. Corpus of Historical American English (COHA). https://doi.org/10.7910/DVN/8SRSYK.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. pages 120–128.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 34, pages 7659–7666.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. pages 1968–1994.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods* 51(3):987–1006. https://link.springer.com/article/10.3758/s13428-018-1115-7.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. pages 1–14.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 6132–6142. https://doi.org/10.18653/v1/D19-1635.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. Robbie: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pages 3764–3814.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 1696–1705.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 6970–6984. https://doi.org/10.18653/v1/2021.acl-long.542.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, pages 1256–1266. https://doi.org/10.18653/v1/2021.eacl-main.107.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 36, pages 11954–11962.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. Comparing intrinsic gender bias evaluation measures without using human annotated examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, pages 2857–2863.

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

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018* page 43.

RJ Kuo and Shu-Syun Li. 2023. Applying particle swarm optimization algorithm-based collaborative filtering recommender system considering rating and review. *Applied Soft Computing* page 110038.

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

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multi-dimensional analysis of biases in Arabic word embeddings. In Imed Zitouni, Muhammad Abdul-Mageed, Houda Bouamor, Fethi Bougares, Mahmoud El-Haj, Nadi Tomeh, and Wajdi Zaghouani, editors, *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Barcelona, Spain (Online), pages 192–199. https://aclanthology.org/2020.wanlp-1.17.

Juanhui Li, Wei Zeng, Suqi Cheng, Yao Ma, Jiliang Tang, Shuaiqiang Wang, and Dawei Yin. 2023. Graph enhanced bert for query understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 3315–3319.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 5502–5515. https://doi.org/10.18653/v1/2020.acl-main.488.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020b. Monolingual and multilingual reduction of gender bias in contextualized representations. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 5082–5093. https://doi.org/10.18653/v1/2020.coling-main.446.

Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. Open-ended long text generation via masked language modeling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 223–241. https://doi.org/10.18653/v1/2023.acl-long.13.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022a. TimeLMs: Diachronic Language Models from Twitter.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022b. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pages 251–260.

Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Razin, Ori Katz, and Noam Koenigstein. 2020. RecoBERT: A catalog language model for text-based recommendations. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, pages 1704–1714. https://doi.org/10.18653/v1/2020.findings-emnlp.154.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 622–628. https://doi.org/10.18653/v1/N19-1063.

Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864* .

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154.

Yeongjoon Park, Youngjoong Ko, and Jungyun Seo. 2022. Bert-based response selection in dialogue sys-

tems using utterance attention mechanisms. *Expert systems with applications* 209:118277.

Wenjun Qiu and Yang Xu. 2022. HistBERT: A Pre-trained Language Model for Diachronic Lexical Semantic Analysis.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 2887–2896. https://doi.org/10.18653/v1/2023.emnlp-main.175.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, pages 1498–1508.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics* 9:845–874.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pages 9180–9211. https://doi.org/10.18653/v1/2022.emnlp-main.625.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 167–177. https://doi.org/10.18653/v1/2021.acl-long.14.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023a. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*. pages 3575–3590.

Xiaohang Tang, Yi Zhou, and Danushka Bollegala. 2023b. Learning dynamic contextualised word embeddings via template-based temporal adaptation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 9352–9369. https://doi.org/10.18653/v1/2023.acl-long.520.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability* 57(1).

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032* .

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 4076–4089.

Ziqian Zeng, Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Socialized word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California. https://www.ijcai.org/proceedings/2017/0547.pdf.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 629–634. https://doi.org/10.18653/v1/N19-1064.

Yi Zhou, Jose Camacho-Collados, and Danushka Bollegala. 2023. A predictive factor analysis of social biases and task-performance in pretrained masked language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 11082–11100. https://doi.org/10.18653/v1/2023.emnlp-main.683.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 1924–1935.

## A Statistics of Temporal Corpora Collected from X

The statistics of temporal corpora collected from X using Twitter's Academic API across a two-year

time span (i.e., from the year 2020 to 2022) can be found in Table 2

| Quarter | 2020 | 2021 | 2022 |
|---|---|---|---|
| Q1 | 7,917,521 | 9,346,385 | 18,708,819 |
| Q2 | 7,922,090 | 9,074,847 | 18,536,812 |
| Q3 | 7,839,401 | 9,388,844 | 18,347,979 |
| Q4 | 7,769,658 | 9,471,075 | 18,427,616 |
| Total | 31,448,670 | 37,281,151 | 74,021,226 |

Table 2: The statistics of temporal corpora collected from X. Each quarter corresponds to three months. Q1: January-March, Q2: April-June, Q3: July-September, Q4: October-December.

## B All Unmasked Likelihood with Attention (AULA)

We compare the pseudo-likelihood scores returned by an MLM for stereotypical and anti-stereotypical sentences using AULA. This metric evaluates social biases by using MLM attention weights to reflect token significance.

Given a sentence $S = s_1, \ldots, s_n$ encompassing a sequence of tokens $s_i$ with a length of $|N|$, we calculate the Pseudo Log-Likelihood, denoted as $\mathrm{PLL}(S)$, to predict all tokens within sentence $S$, excluding the start and end tokens of the sentence. The score $\mathrm{PLL}(S)$ for sentence $S$ given by (2) can be used to assess the preference expressed by an MLM for the given sentence $S$.

$$\mathrm{PLL}(S) := \frac{1}{|N|} \sum_{i=1}^{|N|} \alpha_i \log P(s_i|S;\theta) \quad (2)$$

where $\alpha_i$ is the average of multi-head attention weights associated with each token $s_i$. $P(s_i|S;\theta)$ indicates the probability of the MLM assigning token $s_i$ given the context of sentence $S$. The fraction of sentence pairs where the MLM's preference for stereotypical ($S^{st}$) sentences over anti-stereotypical ($S^{at}$) ones is computed as the AULA bias score of the MLM as in (3).

$$\mathrm{AULA} = \frac{100}{M} \sum_{(S^{\mathrm{st}}, S^{\mathrm{at}})} \mathbb{I}(\mathrm{PLL}(S^{\mathrm{st}}) > \mathrm{PLL}(S^{\mathrm{at}})) \quad (3)$$

Here $M$ denotes the overall count of sentence pairs in the dataset and $\mathbb{I}$ represents the indicator function that yields 1 when its condition is true and 0 otherwise. The AULA score calculated by (3) lies in the interval [0, 100]. An unbiased model would yield bias scores close to 50, while bias scores lower or higher than 50 indicate a bias towards the anti-stereotypical or stereotypical group, respectively.

## C Demographic Descriptor Words for Biases

### C.1 Gender Bias

The names associated with female and male for gender biases are listed in Table 3.

### C.2 Race Bias

The names associated with two different demographic groups for race bias are listed in Table 4.

### C.3 Religion Bias

The terms associated with two different demographic groups for religion bias are listed in Table 5.

### C.4 Age Bias

The terms associated with two different demographic groups for religion bias are listed in Table 6.

## D Social bias of the control experiment

Table 7 and Table 8 show the social bias scores across time on the temporal corpora collected from X and the 5 subsets of corpus randomly sampled from a fixed time period, respectively.

Table 9 shows the standard deviation of social biases with different timestamps and within the same periods.

## E Results of COHABERT

### E.1 Biases in COHABERT

The result of the bias scores computed on both CrowS-Pairs and StereoSet for different bias types in COHABERT is shown in Figure 4. The average and standard deviations are computed based on the AULA bias scores covering a period of 190 years, specifically from 1810 to 2000, with scores provided for each decade.

### E.2 Statistical Indicators of Bias Fluctuation Changes in COHABERT

The statistical indicators of bias fluctuation changes in COHABERT models are shown in Table 10.

| Demographic Group | Terms |
|---|---|
| Female | Olivia, Emma, Charlotte, Amelia, Sophia, Isabella, Ava, Mia, Evelyn, Luna, Harper, Camila, Sofia, Scarlett, Elizabeth, Eleanor, Emily, Chloe, Mila, Violet, Penelope, Gianna, Aria, Abigail, Ella, Avery, Hazel, Nora, Layla, Lily, Aurora, Nova, Ellie, Madison, Grace, Isla, Willow, Zoe, Riley, Stella, Eliana, Ivy, Victoria, Emilia, Zoey, Naomi, Hannah, Lucy, Elena, Lillian |
| Male | Liam, Noah, Oliver, James, Elijah, William, Henry, Lucas, Benjamin, Theodore, Mateo, Levi, Sebastian, Daniel, Jack, Michael, Alexander, Owen, Asher, Samuel, Ethan, Leo, Jackson, Mason, Ezra, John, Hudson, Luca, Aiden, Joseph, David, Jacob, Logan, Luke, Julian, Gabriel, Grayson, Wyatt, Matthew, Maverick, Dylan, Isaac, Elias, Anthony, Thomas, Jayden, Carter, Santiago, Ezekiel, Charles |

Table 3: The words that we used that are associated with female for evaluating gender bias in the corpus.

| Demographic Group | Terms |
|---|---|
| African American | Ebony, Jasmine, Lakisha, Latisha, Latoya, Nichelle, Shaniqua, Shereen, Tanisha, Tia, Alonzo, Alphonse, Darnell, Jamel, Jerome, Lamar, Leroy, Malik, Terrence, Torrance |
| European American | Amanda, Betsy, Courtney, Ellen, Heather, Katie, Kristin, Melanie, Nancy, Stephanie, Adam, Alan, Andrew, Frank, Harry, Jack, Josh, Justin, Roger, Ryan |

Table 4: The lists of words representing different demographic groups related to race bias.

| Demographic Group | Terms |
|---|---|
| Christian | christianize, christianese, Christians, christian-only, christianising, christiansand, christiany, jewish-christian, -christian, Christian., christianise, christianists, Christian, Christianity, christian-, Christians., christianity-, Christianity., christian-muslim, muslim-christian, christianized, christian-right, christianist, christian-jewish |
| Jewish | judaïsme, jewish-canadian, half-jewish, part-jewish, anglo-jewish, jewes, french-jewish, -jewish, jewish-related, jewsish, christian-jewish, jewish-, jewish-zionist, anti-jewish, jewish-muslim, jewishgen, jews-, jewish-american, jewish., jewish-roman, jewish-german, jewish-christian, jewishness, american-jewish, jewsih, jewish-americans, jewish-catholic, jewish, jew-ish, spanish-jewish, semitic, black-jewish, jewish-palestinian, jewish-christians, jew, jewish-arab, jews, russian-jewish, jewish-owned, jew., german-jewish, judaism, jewishly, muslim-jewish, judaism., jewish-italian, jewish-born, all-jewish, austrian-jewish, catholic-jewish, jews., judaism-related, roman-jewish, jewish-themed, college-jewish, arab-jewish, jewish-only, british-jewish, judaisms, jewish-russian, pro-jewish, israeli-jewish, jewish-israeli |

Table 5: The lists of words representing different demographic groups related to religion bias.

| Demographic Group | Terms |
|---|---|
| young | adolescent, teen, teenage, teenaged, young, younger, twenty-year-old, 20-year-old, twentyfive-year-old, 25-year-old, thirty-year-old, 30-year-old, thirty-five-year-old, 35-year-old, forty-year-old, 40-year-old, twenty-something, thirty-something |
| old | sixty-five-year-old, 65-year-old, seventy-year-old, 70-year-old, seventy-five-year-old, 75-year-old, eighty-year-old, 80-year-old, eighty-five-year-old, 85-year-old, ninety-year-old, 90-year-old, ninety-five-year-old, 95-year-old, seventy-something, eighty-something, ninety-something, octogenarian, nonagenarian, centenarian, older, old, elderly, retired, senior, seniorcitizen, young-at-heart, spry |

Table 6: The lists of words representing different demographic groups related to religion bias.

| Bias Scores | Female bias | Male bias |
|---|---|---|
| Mar 2020 | 62.05 | 30.17 |
| Jun 2020 | 64.01 | 31.01 |
| Sep 2020 | 63.53 | 31.44 |
| Dec 2020 | 61.90 | 31.28 |
| Mar 2021 | 60.79 | 30.97 |
| Jun 2021 | 57.96 | 29.83 |
| Sep 2021 | 61.45 | 30.24 |
| Dec 2021 | 58.64 | 30.55 |
| Mar 2022 | 59.76 | 31.74 |
| Jun 2022 | 62.51 | 32.65 |
| Sep 2022 | 63.77 | 31.84 |

Table 7: The social bias score of temporal corpora collected from X.

| Bias Scores | Female bias | Male bias |
|---|---|---|
| sample 1 | 62.15 | 59.89 |
| sample 2 | 62.36 | 60.34 |
| sample 3 | 61.99 | 60.19 |
| sample 4 | 62.36 | 60.21 |
| sample 5 | 62.18 | 59.96 |

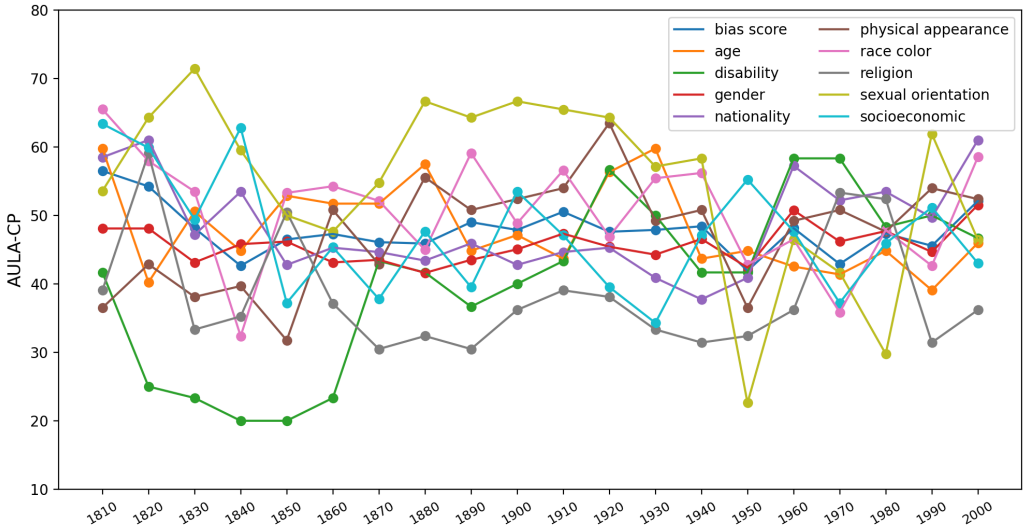
Table 8: The social bias score of 5 subsets of corpus randomly sampled from Jan to Mar 2020.

| Standard deviation | Female bias | Male bias |
|---|---|---|
| across time | 2.03 | 0.84 |
| same timestamp | 0.16 | 0.19 |

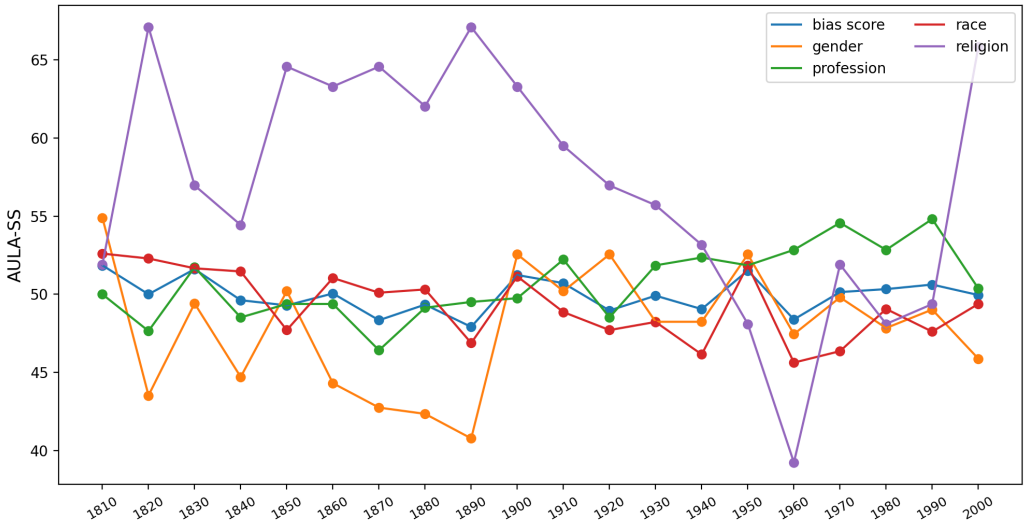
Table 9: The standard deviations of temporal corpora collected from X and the subset of corpus random sampled from January to March 2020.

| | mean | lower/upper | SE | SD |
|---|---|---|---|---|
| **CrowS-Pairs** | | | | |
| OVERALL BIAS | 47.83 | 46.59/49.23 | 0.79 | 3.59 |
| sexual-orientation | 54.64 | 49.35/58.63 | 2.77 | 12.74 |
| disability | 40.50 | 35.92/44.68 | 2.68 | 12.32 |
| socioeconomic | 47.24 | 44.39/50.52 | 1.85 | 8.54 |
| religion | 38.38 | 35.81/42.05 | 1.83 | 8.47 |
| race-color | 50.56 | 47.40/53.27 | 1.77 | 8.21 |
| appearance | 47.46 | 44.52/50.16 | 1.73 | 7.89 |
| nationality | 48.40 | 46.04/51.16 | 1.54 | 7.02 |
| age | 48.16 | 46.04/50.69 | 1.40 | 6.45 |
| gender | 45.74 | 44.86/46.75 | 0.58 | 2.65 |
| **StereoSet** | | | | |
| OVERALL BIAS | 49.94 | 49.54/50.34 | 0.24 | 1.11 |
| religion | 57.15 | 54.18/59.75 | 1.68 | 7.65 |
| gender | 47.86 | 46.49/49.27 | 0.85 | 3.88 |
| profession | 50.69 | 49.89/51.51 | 0.49 | 2.24 |
| race | 49.30 | 48.51/50.08 | 0.48 | 2.20 |

Table 10: The confidence interval and standard error computed using bootstrapping for each of the bias types on the CrowS-Pairs and StereoSet benchmarks for COHABERT models. SE and SD represent standard error and standard deviation, respectively. Lower/upper indicates the lower/upper bound of the confidence intervals. In each dataset, different bias types are sorted in the descending order of their SD.




(a) CrowS-Pairs




(b) StereoSet

Figure 4: Social bias scores across time for different types of biases computed using the AULA metric for COHABERT models. Results evaluated on the CrowS-Pairs and StereoSet datasets are shown respectively on the top and bottom. The 'bias score' (in dark blue) indicates the overall bias score.