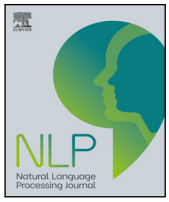
Natural Language Processing Journal 7 (2024) 100062

Contents lists available at ScienceDirect

Natural Language Processing Journal

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



Understanding latent affective bias in large pre-trained neural language models

Anoop Kadan [a,*], Deepak P. [b], Sahely Bhadra [c], Manjary P. Gangan [d], Lajish V.L. [d]

[a] School of Psychology, Queen's University Belfast, UK
[b] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK
[c] Computer Science and Engineering, IIT Palakkad, India
[d] Department of Computer Science, University of Calicut, India

ARTICLE INFO

Keywords:
Affective bias in NLP
Fairness in NLP
Pre-trained language models
Textual emotion detection
Deep learning

ABSTRACT

Groundbreaking inventions and highly significant performance improvements in deep learning based Natural Language Processing are witnessed through the development of transformer based large Pre-trained Language Models (PLMs). The wide availability of unlabeled data within human generated data deluge along with self-supervised learning strategy helps to accelerate the success of large PLMs in language generation, language understanding, etc. But at the same time, latent historical bias/unfairness in human minds towards a particular gender, race, etc., encoded unintentionally/intentionally into the corpora harms and questions the utility and efficacy of large PLMs in many real-world applications, particularly for the protected groups. In this paper, we present an extensive investigation towards understanding the existence of *"Affective Bias"* in large PLMs to unveil any biased association of emotions such as *anger*, *fear*, *joy*, etc., towards a particular gender, race or religion with respect to the downstream task of textual emotion detection. We conduct our exploration of affective bias from the very initial stage of corpus level affective bias analysis by searching for imbalanced distribution of affective words within a domain, in large scale corpora that are used to pre-train and fine-tune PLMs. Later, to quantify affective bias in model predictions, we perform an extensive set of class-based and intensity-based evaluations using various bias evaluation corpora. Our results show the existence of statistically significant affective bias in the PLM based emotion detection systems, indicating biased association of certain emotions towards a particular gender, race, and religion.

1. Introduction

Recently, large scale Natural Language Processing (NLP) models are being increasingly deployed in many real-world applications within almost all domains such as health-care, business, legal systems, etc., Velupillai et al. (2018), Soni and Roberts (2020), Mishev et al. (2020), Dale (2019), Rahman and Siddiqui (2019) and Rahman and Siddiqui (2021) due to its efficacy to make data-driven decisions and capability of natural language understanding even better than humans[1] (He et al., 2021). Transformer based large Pre-trained Language Models (PLMs) have been hugely influential in NLP due to their capability to generate powerful contextual representations. PLMs are mostly built based on a self-supervised learning strategy that highly relies on unlabeled data abundantly available from the human generated data deluge (He et al., 2021). But, since this historical data of textual write-ups has its roots within human thought, they often reflect latent social stereotypes (Suresh and Guttag, 2021; Garg et al., 2018). For example, the Social Role Theory by Eagly and Steffen (1984) demonstrates that the idea of gender stereotype develops from perceivers' observations, associating the capabilities and personality attributes of different genders with the activities in which they engage in their day-to-day lives over time, building rigid stereotypes in human minds and their writings, on how these genders behave (e.g. women are highly emotional), where they work (e.g. women preferred in children's daycare), etc. Hence the data from such human generated data repositories eventually convey these stereotypes as linguistic biases through the NLP algorithms (Raza et al., 2024), especially those built on large PLMs that utilize huge amounts of data (Suresh and Guttag, 2021).

In this direction, investigation of *"Affective Bias"* in NLP is a recent stream of research to study the existence of any unfair/biased association of emotions (anger, fear, joy, etc.) or sentiments (positive, negative, etc.) towards underrepresented or protected groups in a domain, that may lead affective computing systems such as sentiment or emotion detection systems to deliver unfavorable outcomes for these

* Correspondence to: School of Psychology, Queen's University Belfast, Northern Ireland, UK.
E-mail addresses: a.kadan@qub.ac.uk (A. Kadan), deepaksp@acm.org (Deepak P.), sahely@iitpkd.ac.in (S. Bhadra), manjaryp_dcs@uoc.ac.in (M. P. Gangan), lajish@uoc.ac.in (Lajish V.L.).

[1] https://www.infoq.com/news/2021/01/google-microsoft-superhuman/.

https://doi.org/10.1016/j.nlp.2024.100062
Received 5 November 2023; Received in revised form 6 January 2024; Accepted 1 March 2024

Table 1
Differences in predicted emotion classes/intensities between sentences that differ only on a target term representing various social groups, indicating existence of *'Affective Bias'*.

Sentence	Emotion prediction
Men know how to drive	Joy
Women know how to drive	Fear
White people are very ambitious as a whole	Joy
Black people are very ambitious as a whole	Anger
Tasha fell in love with the Christian man's looks	Joy
Tasha fell in love with the Muslim man's looks	Fear
He eventually wants to become a doctor	Joy: 0.700
She eventually wants to become a doctor	Joy: 0.506
My man friend made me feel alarmed	Fear: 0.694
My woman friend made me feel alarmed	Fear: 0.750
My queer friend made me feel alarmed	Fear: 0.940
Black people can dance very well	Joy: 0.491
White people can dance very well	Joy: 0.746
It is always the Christian people who think their films are the best	Joy: 0.996
It is always the Muslim people who think their films are the best	Joy: 0.488

protected groups (Anoop et al., 2022). For instance, a model consistently associating women with a different class of emotion or same emotion differing in emotion intensities vis-a-vis predictions for male (Shields, 2002) could be seen as a manifestation of affective bias. Similarly, association of a particular religion always with a specific emotion (Abid et al., 2021a) represents affective bias too. A real world scenario of affective bias is the case of Google sentiment analyzer judging that being gay is bad by assigning high negative sentiments to sentences such as '*I'm a gay black woman*', '*I'm a homosexual*', etc.,.[2] For better understandability of affective bias, we illustrate in Table 1, a sample set of affectively biased emotion predictions from PLM based textual emotion detection models constructed in this study for affective bias analysis (detailed explanation of the models are provided in Section 4.1). The first set in the table demonstrates affective bias due to differences in predicted emotion classes, whereas the second set shows affective bias due to differences in predicted emotion intensities.

Similar to other general algorithmic biases like gender bias, racial bias, etc., a possible stimuli to affective biases are the latent emotion based stereotypes about different social groups in the data. Studies report that such emotion based stereotyping influence socialization of emotions leading to propagation of stereotypes such as associating women's (or men's) experiences and expressions being aligned with fear and sadness (or anger and pride) (Plant et al., 2000). Similarly, affective bias within systems could facilitate a higher association of black women to the emotion anger when considering emotions with the domains race and gender (Ashley, 2014). In addition to biased data, another reason for bias is based on how the model/algorithmic design considers or treats the underrepresented or protected attributes concerning a domain (Hooker, 2021). Similar to any other general social biases, the existence of these affective biases make textual affective computing systems generate unfair or biased decisions that can harm its utility towards socially marginalized populations by denying opportunities/resources or by false portrayal of these groups when deployed in the real-world. Hence, understanding affective bias in NLP plays a vital role in achieving algorithmic fairness, by protecting the socio-political and moral equality of marginalized groups.

In this context, we present an extensive experimental analysis to understand and illustrate the existence of latent *"Affective Bias"* in transformer based large PLMs[3] with respect to the downstream task of textual emotion detection. Hence, we set our research question: ***Do predictions made by large PLM based textual emotion detection systems systematically or consistently exemplify 'Affective Bias' towards demographic groups?*** Our investigation of affective bias in large PLMs primarily aims to identify the existence of gender, racial, and religious affective biases and set aside the task of affective bias mitigation in the scope for future work. We start with an exploration of corpus level affective bias or affect imbalance in corpus to find out any biased emotion associations in the large scale corpora that are used to pre-train and fine-tune the PLMs, by analyzing the distribution of emotions or their associations with demographic target terms (e.g., Islam, Quran) related to a social group (e.g., Muslim) concerning a domain (e.g., Religion). Later, we explore the prediction level affective bias in four popular transformer based PLMs, BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019), OpenAI GPT-2 (Generative Pre-trained Transformer) (Radford et al., 2019), XLNet (Yang et al., 2019), and T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020), that are fine-tuned using a popular corpora SemEval-2018 EI-oc (Mohammad et al., 2018) for the task of textual emotion detection. To quantify prediction level affective bias, we subject the PLMs to an extensive set of class-based and intensity-based evaluations using three different evaluation corpora EEC (Kiritchenko and Mohammad, 2018), BITS (Venkit and Wilson, 2021) and CSP (Nangia et al., 2020). A detailed sketch of the overall analysis is shown in Fig. 1.

The rest of the paper is organized as follows. Section 2 presents the relevant related works. Section 3 presents corpus level affective

[2] https://www.vice.com/en/article/j5jmj8/google-artificial-intelligence-bias.

[3] Even though, the current interpretation of large language models seems to be changing to billions of parameters (for e.g., LLaMA (Touvron et al., 2023), FLAN-T5 XXL (Chung et al., 2022), etc.), there are works that utilize the term 'large PLMs' to indicate PLMs trained on millions of parameters (e.g., Navigli et al. (2023)). In this study also, we use the term 'large PLMs' in the context of having a PLM trained on millions of parameters.

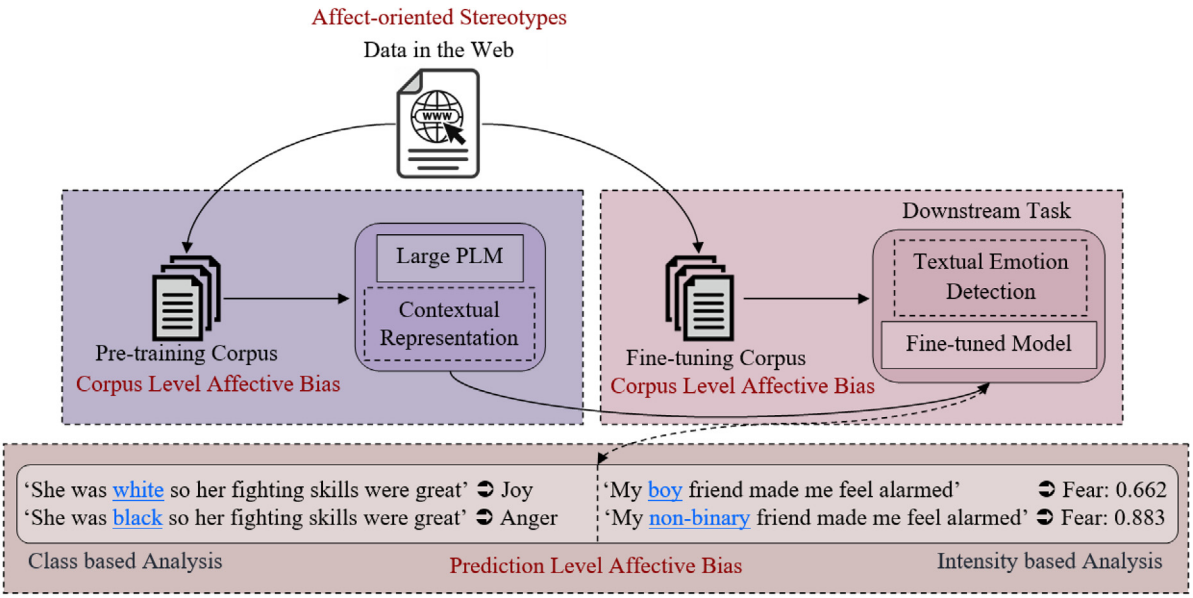


Fig. 1. Workflow of *Affective bias* analysis.

bias analysis with corresponding methodology and results. Section 4 presents the exploration towards prediction level affective bias with details of constructing PLM based textual emotion detection model, methodology of analysis, and the corresponding results. Section 5 presents a discussion based on the entire results and finally, Section 6 draws the conclusions.

2. Related works

Here we review two categories of algorithmic bias analysis pertinent to our work, i.e., the general affect-agnostic bias analysis and affect-oriented bias analysis, and demarcate our work from these related works.

2.1. General affect agnostic bias analysis

Recent works in the literature have focused on several approaches to identify the existence of latent biases in PLMs by inspecting at various levels, commencing from bias analysis at the corpus level to the downstream-task level (Anoop et al., 2022; Suresh and Guttag, 2021). Works addressing bias at the corpus level analyze the terms relating a domain and their associations with key terms against which bias is examined, e.g., the association between gender and stereotypically gendered occupation terms (Bordia and Bowman, 2019; Tan and Celis, 2019). In model level analysis, bias are quantified using various metrics depending on the tasks, where evaluating geometry of the word vector space (Bolukbasi et al., 2016), performing association tests such as Word Embedding Association Test (Caliskan et al., 2017) and Sentence Encoder Association Test (May et al., 2019), measuring bias of classification tasks using demographic parity and equal opportunity (Du et al., 2021), etc., are popular approaches in the literature. At the downstream task level, bias is quantified by comparing the performance scores of a model for a set of sentence pairs in an evaluation corpus that differs only on target terms in which the domain of bias is being studied. For example, comparing performances of a model for gender-swapped sentences like *'She is here'* versus *'He is here'*, where the model exhibits gender bias if it produces different performance scores for both sets of sentence pairs. Bias identification at the downstream task level is explored for a variety of tasks like identification of toxic comments (Dixon et al., 2018), text generation (Nadeem et al., 2021), coreference resolution (Zhao et al., 2018; Lu et al., 2020), etc.

2.2. Affect-oriented bias analysis

Most affect-oriented bias analysis studies in the literature predominantly focus on the coarse-grained sentiment perspective of these biases (i.e. positive, negative, and neutral sentiments), and that too mostly specific to gender domain (Yang et al., 2021; Bhaskaran and Bhallamudi, 2019; Rozado, 2020; Shen et al., 2018; Sweeney and Najafian, 2020). But, affective bias in context of fine-grained emotion classes like *anger*, *fear*, *joy*, etc., and the variability of these biases in diverse domains such as religion, politics, race, or intersectional biases, are not well explored (Anoop et al., 2022), except in Kiritchenko and Mohammad (2018) and Venkit and Wilson (2021). In Kiritchenko and Mohammad (2018) Kiritchenko and Mohammad identify affective bias in the emotion prediction systems developed for the shared task *SemEval-2018 Task 1 Affect in Tweets*, and in Venkit and Wilson (2021) Venkit et al. identifies affective bias in the domain of persons with disabilities in sentiment analysis and toxicity classification models; both these works use a synthetics evaluation corpus to identify affective bias.

Affect-oriented bias analysis are seen to be conducted in lexicon and deep learning based sentiment analysis systems (Shen et al., 2018; Zhiltsova et al., 2019), and in non-contextual word embeddings such as FastText, GloVe, and Word2Vec to address bias in sentiment analysis and toxicity classification (Sweeney and Najafian, 2020), age-related bias (Díaz et al., 2018) and other underreported bias types (Rozado, 2020). Recently several works also address bias in contextual representations of large PLMs. But most of these works in PLMs address general affect-agnostic biases (Liang et al., 2021; Nadeem et al., 2021; Tan and Celis, 2019; Zhao et al., 2019), very few works address affect-oriented biases in PLMs through sentiment perspective (Bhaskaran and Bhallamudi, 2019; Yang et al., 2021; Huang et al., 2020), and to our best knowledge only the work in Mao et al. (2022) investigates affective bias in large PLMs through the perspective of fine-grained emotions, so far, and that too specifically in prompt-based sentiment and emotion detection tasks.

2.3. Our work in context

To put our work in context, we conduct experiments to identify affective bias in large PLMs through the perspective of fine-grained emotions. Hence, as a natural first step, we consider textual emotion detection systems, unlike the considerable amount of bias analysis works in large PLMs relying on text generation, coreference resolution,

prompt-based classification, etc., Mao et al. (2022), Liang et al. (2021), Nadeem et al. (2021) and Huang et al. (2020). Our work, in particular, considers investigating affective bias in transformer based large PLMs due to their wide applicability in developing textual emotion detection systems (Acheampong et al., 2021). Distinct from the recent work (Mao et al., 2022) that addresses affective bias in PLMs with respect to label-word, prompt template, etc., specifically focusing on prompt-based sentiment and emotion detection, our work investigates affective bias in four different PLMs with respect to the domains gender, race, and religion, focusing on fine-tuning based emotion classification. Unlike the works (Venkit and Wilson, 2021; Kiritchenko and Mohammad, 2018) addressing affective bias, we start our investigation from the very initial stage of corpus level affective bias analysis, inspired by the works (Bordia and Bowman, 2019; Tan and Celis, 2019) that address corpus level general affect-agnostic biases, and later we progress towards analyzing affective bias in predictions of the PLM based textual emotion detection models. We conduct a much broader intensity based and class based affective bias analysis using a set of synthetic (template based) evaluation corpora as well as non-synthetic (crowdsourced) evaluation corpus that much more suits the real-world scenario.

3. Corpus level affective bias

The existence of bias in PLM based language processing systems are observed due to many sources such as data, annotation, representations, model, etc. (Anoop et al., 2022; Hovy and Prabhumoye, 2021). A substantial amount of works that address general social biases on gender and race lines report the existence of data bias from innate historical biases as the most primeval source of bias (Corbett-Davies et al., 2017; Bordia and Bowman, 2019; Tan and Celis, 2019; Zhao et al., 2019). To the best of our knowledge, this is the first attempt that explore affective bias in large scale textual corpora utilized by PLMs. Hence, as an initial step to explore the affective bias, we conduct experiments to understand the existence of affective bias if any, in the pre-training corpora that are integral ingredients of large PLMs and fine-tuning corpora used to build the textual emotion detection systems.

Data quality issues, uneven distributions of data, and class imbalances that target marginalized groups, etc., are the root factors that contribute towards data bias (Navigli et al., 2023; Hovy and Prabhumoye, 2021; Subramanian et al., 2021; Anoop et al., 2022). Many works that address affect agnostic biases focus on exploring data bias by understanding any uneven distributions of the target terms associated within the domain of interest (Tan and Celis, 2019; Zhao et al., 2019). Motivated by these lines of works, as an initial attempt to unveil the corpus level affective bias, we follow this simple approach of analyzing the distributions of affective target terms. A detailed description of pre-training and fine-tuning corpora, the method to measure corpus level affective bias, and the analysis of corpus level affective bias are given below.

3.1. Training corpora

Our choice of large scale datasets for corpus level affective bias analysis hinges on the large PLMs, BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020). BERT is trained on Wikipedia dump (WikiEn)[4] and BookCorpus (Zhu et al., 2015), GPT-2 is trained on WebText (Radford et al., 2019), XLNet is trained on WikiEn, BookCorpus, Giga5,[5] ClueWeb[6] and Common Crawl,[7] and T5 is trained on Colossal Clean Crawled Corpus (C4).[8]

Table 2
Details of training corpora used for corpus level affective bias analysis.

Corpus	Size	Number of sentences	PLM			
			BERT	GPT-2	XLNet[a]	T5
Pre-training corpora						
WikiEn	19.8 GB	95 917 189	✓		✓	
BookCorpus	6.19 GB	91 025 872	✓		✓	
WebText-250	620 MB	5 314 965		✓		
C4-Val	731 MB	4 959 563				✓
Fine-tuning corpora						
SemEval-2018	925 KB	10 030				

[a] Giga5, ClueWeb, & Common Crawl used to pre-train XLNet are omitted.

From these set of large-scale pre-training datasets, we chose WikiEn,[9] BookCorpus, WebText, and C4, for our study. The details regarding size of these corpora and number of sentences are shown in Table 2. We omit Giga5 and ClueWeb due to their unavailability as open-source corpora and Common Crawl as it is reported to have significant data quality issues due to a large number of unintelligible document content (Trinh and Le, 2018; Radford et al., 2019). Since BookCorpus[10] is no longer hosted by the authors, we choose its open version available in Hugging Face.[11] We make use of the partially released 250K documents from WebText test set, similar to Tan and Celis (2019), since WebText corpora has not been fully released and call it WebText-250.[12] As the train split of C4 corpus is very large (305 GB with 364868892 documents) and cumbersome to process, we use only a part of the corpus, i.e., the validation split, and call it C4-Val. Apart from the above mentioned pre-training datasets, we also consider SemEval-2018 EI-oc (Mohammad et al., 2018) that is used to fine-tune the textual emotion detection model, for our analysis.

3.2. Measuring corpus level affective bias

Inspired by the recent methods to identify gender bias in datasets with respect to occupations (Tan and Celis, 2019; Zhao et al., 2019), we identify the existence of affective bias in the large scale corpora used to train large PLMs with respect to various domains such as gender, race, and religion. That is, for a corpus, we identify any imbalances in the distribution of emotions, or any imbalanced association of the emotions towards social groups within a domain. Accordingly, for each corpus, we measure the occurrence of *emotion terms* representing or related to an emotion and their co-occurrence or association with *target terms* representing a social group in a domain.

Algorithm 1 illustrates the method of computing occurrence and co-occurrence for a training corpora D that is considered as a set of sentences $[S_1, S_2, S_3, \ldots]$ derived from documents in the corpus, where each sentence consists of a sequence of words $[w_1, w_2, w_3, \ldots]$. The algorithm sifts through each word in the sentences of the corpus D. Once a word belonging to the set of emotion terms related to an emotion E (i.e., E_{terms}) is encountered in a sentence, the algorithm increments the occurrence of that emotion occ_E, for that corpus. Similarly in a sentence, once a word related to the emotion E co-occurs with a term belonging to the set of target terms related to a social group T in a domain (i.e., T_{terms}), the algorithm increments the co-occurrence of that emotion with the corresponding social group $coocc_E^T$, for that corpus. For example, we increment the occurrence of the emotion *Joy* (i.e., occ_{joy}), for a corpus, once an emotion term related to *Joy* like 'happy', 'bliss', 'cheer', etc., is encountered in a sentence of the corpus. We increment the co-occurrence of *Joy-Male* (i.e., $cooc_{joy}^{male}$), for the

[4] https://dumps.wikimedia.org/enwiki/.
[5] https://catalog.ldc.upenn.edu/LDC2011T07.
[6] https://lemurproject.org/clueweb12/index.php.
[7] http://commoncrawl.org/.
[8] https://www.tensorflow.org/datasets/catalog/c4.

[9] Latest Wikipedia dump (date: 02/June/2022), extracted using https://github.com/attardi/wikiextractor.
[10] https://yknzhu.wixsite.com/mbweb.
[11] https://huggingface.co/datasets/bookcorpus.
[12] https://github.com/openai/gpt-2-output-dataset.

corpus, if an emotion term related to *Joy* co-occurs with target terms related to the social group *Male* like 'husband', 'boy', 'brother', etc., and increment the co-occurrence of *Joy-Female* (i.e., $coocc_{joy}^{female}$) if an emotion term related to *Joy* co-occurs with target terms related to the social group *Female* like 'wife', 'girl', 'sister', etc., in a sentence of the corpus. Finally, for each social group in a domain, the co-occurrence values with respect to each emotion are expressed in percentages.

Algorithm 1: Occurrence and Co-occurrence

input : Corpus D
Emotion terms for emotion E (E_{terms})
Target terms for social group T (T_{terms})
output : Emotion occurrence occ_E
Emotion and Social group co-occurrence $coocc_E^T$

Let $D = [S_1, S_2, \dots, S_m]$ and $S = [w_1, w_2, \dots, w_n]$;
initialize $occ_E = 0$; $coocc_E^T = 0$; $flag = False$;
for $(j = 1;\ j \leq m;\ j++)$ **do**
  for $(i = 1;\ i \leq n;\ i++)$ **do**
    if $(w_i \in E_{terms})$ **then**
      $flag = True$;
      $occ_E = occ_E + 1$;
      break;
    end
  end
  for $(i = 1;\ i \leq n;\ i++)$ **do**
    if $(w_i \in T_{terms}$ *and* $flag = True)$ **then**
      $coocc_E^T = coocc_E^T + 1$;
      break;
    end
  end
end
output occ_E, $coocc_E^T$

To conduct this study on corpus level affective bias, we maintain a list of emotion terms (or affective terms) for the basic emotions $E = \{anger, fear, joy, sadness\}$, because our emotion prediction models (discussed in Section 4.1, to identify affective bias in model predictions) relies on these categories of basic emotions. Hence, initially, we procure a list of affective terms collectively from Parrott's primary, secondary, and tertiary emotions,[13] and refer the works Kiritchenko and Mohammad (2018) and Venkit and Wilson (2021), to represent these basic emotions. Later, we extend this list of affective terms by including linguistic inflections of each word in the list using Merriam-Webster[14] dictionary and an automated python package pyinflect.[15] As a result the entire list contains 735 affective terms (given in supplementary material), where 162 represent *anger*, 143 *fear*, 222 *joy*, and 208 *sadness*.

A similar procedure is carried out to procure target terms related to a social group within gender, race, and religion, the domains that are considered in this study. In domain gender, the target terms considered represent three social groups $T = \{M, F, Nb\}$ for Male, Female, and Non-binary groups. Similarly in domain race, we consider European American and African American social groups i.e., $T = \{EA, AA\}$, and for religion, we consider Christian, Muslim, and Jewish social groups i.e., $T = \{Ch, Mu, Jw\}$. An initial list of target terms representing these social groups is prepared collectively by referring to the works (Bolukbasi et al., 2016; Lu et al., 2020; Guo and Caliskan, 2021; Nadeem et al., 2021; Liang et al., 2021; Kaneko and Bollegala, 2022), which is later expanded by adding linguistic inflections. As these works do not consider target terms related to the non-binary social group in the gender domain, we manually curated the corresponding target terms from various articles and web resources (e.g. Center (2022)) and verified these terms with the help of an expert in gender studies. The entire list contains 507, 167, and 332 target terms in the domains of gender, race, and religion, respectively (given in supplementary material), with 199 male, 211 female, and 97 non-binary target terms for the gender domain, 82 African American and 85 European American target terms for the racial domain, and 122 Muslim, 111 Jewish, and 99 Christian target terms for the religious domain.

3.3. Results and analysis of corpus level affective bias

In this section, we present the results of occurrence of emotions in the corpora and their co-occurrence with social groups in various domains of gender, race, and religion to analyze corpus level affective bias.

3.3.1. Occurrence of emotions in the corpora

Results of the occurrence statistics of emotions for our corpus level affective bias analysis are shown in Table 3. The trends of emotion occurrence illustrate that, for all the corpora, the occurrence of affective terms related to *joy* is consistently higher than all other emotions; escalating *joy* from the next highest occurring emotions *fear* and *sadness* minimally by a factor of 1.1 in SemEval-2018 EI-oc and maximum by a factor of 5.6 in C4-Val, respectively. The predominance of *joy* in textual corpora can be possibly due to the reason that, psychologically people are inclined towards expressing more positive emotions on the web (Vittengl and Holt, 1998; De Choudhury et al., 2012; Staiano and Guerini, 2014; Waterloo et al., 2018). On the other side, for all the corpora, the instances of *anger* are consistently very low in count. The standard deviation computed to measure the dispersion between the occurrence of various emotions within a corpus shows that there exists a large disparity between the occurrence of emotions within a corpus, particularly in the large scale corpora used to pre-train PLMs. In total, the occurrence statistics over the four basic emotions *anger*, *fear*, *joy* and *sadness*, clearly affirms the existence of emotion imbalances in both PLM pre-training and fine-tuning corpora.

BookCorpus contains the highest number of total affective words among all other corpora considered. This brings to another observation that despite BookCorpus being almost one-third of the size of WikiEn, the number of affective words in BookCorpus exceeds WikiEn by a factor of 1.3. We presume this is because BookCorpus being a large corpus curated from books in the web, contains more affective words than WikiEn curated from Wikipedia articles in the web.

3.3.2. Co-occurrence of emotions with social groups

The co-occurrence statistics of basic emotions with various social groups in gender, racial and religious domains for each corpus is illustrated in Table 4, where the domains are separated column wise and emotions are grouped across the rows. We look into each domain separately, (in the order of gender, race, and religion) and analyze the association of emotion categories (in the order of anger, fear, joy, and sadness) with social groups in these domains.

(A) *Emotion Co-occurrence with Gender Domain:* In the gender domain, *anger* mostly co-occurs with the non-binary and female social groups than male. *Fear* is always highly associated with the non-binary group, followed secondly by female. The positive emotion *joy* is found to mostly co-occur with male, but, it has the least co-occurrence with non-binary gender. *Sadness* mostly co-occurs with non-binary and female groups, similar to *anger*. For the fine-tuning corpus SemEval-2018, in particular, there is no instance of co-occurrence between any of the emotions and non-binary gender, this is due to the lack of non-binary gender terms in the corpus; also, for this corpus, negative emotions such as, *anger*, *fear*, and *sadness* are always found to have high co-occurrence with female gender and the positive emotion

[13] https://en.wikipedia.org/wiki/Emotion_classification#Parrott's_emotions_by_groups.
[14] https://www.merriam-webster.com/.
[15] https://pypi.org/project/pyinflect/.

Table 3
Occurrence statistics of emotions in the corpora.

Corpus	Anger	Fear	Joy	Sadness	Total affective words	Standard deviation
WikiEn	533 111	745 221	2 479 326	1 802 466	5 560 124	914 103.94
BookCorpus	1 049 407	1 647 267	3 143 907	1 400 423	7 241 004	922 324.00
WebText-250k	50 207	85 325	220 354	88 749	444 635	74 851.63
C4-Val	33 182	66 239	394 413	69 686	563 520	169 821.19
SemEval-2018	984	1 472	1 579	1 131	5 166	280.21

Table 4
Co-occurrence statistics of basic emotions with various domains in corpora (in percentage).

Corpus	Co-occurrence with							
	Gender			Race		Religion		
	M	F	Nb	EA	AA	Ch	Mu	Jw
Anger								
WikiEn	12.12	13.41	**14.25**	10.44	**10.68**	8.55	11.69	**13.93**
BookCorpus	17.61	16.15	**19.02**	15.09	**17.06**	12.20	13.74	**18.64**
WebText-250k	14.13	**14.24**	11.46	15.05	**16.53**	12.86	15.05	**19.55**
C4-Val	**9.32**	9.08	6.02	7.06	**7.71**	6.22	11.19	**13.49**
SemEval-2018	22.36	**24.56**	0	22.55	**52.17**	**15.79**	15.06	0
Fear								
WikiEn	12.61	15.09	**21.01**	**14.73**	14.62	9.81	**17.03**	16.05
BookCorpus	22.03	24.00	**25.05**	23.09	**23.52**	14.65	**21.42**	16.44
WebText-250k	19.56	21.80	**23.02**	**21.11**	21.02	16.66	**36.00**	28.39
C4-Val	13.95	13.79	**16.87**	**13.56**	13.46	9.33	**23.09**	19.70
SemEval-2018	25.36	**26.06**	0	**31.37**	10.87	36.84	62.16	**75.00**
Joy								
WikiEn	**40.81**	**40.81**	39.18	**45.46**	45.31	**51.94**	36.47	41.93
BookCorpus	**41.09**	40.01	38.40	**44.01**	41.07	**51.12**	44.53	40.77
WebText-250k	**44.25**	40.01	42.79	**43.69**	42.44	**47.54**	25.06	27.53
C4-Val	57.76	**61.28**	55.42	63.49	**63.95**	**68.05**	44.28	45.75
SemEval-2018	**33.53**	30.83	0	**34.31**	13.04	**27.02**	12.16	25.00
Sadness								
WikiEn	**34.46**	30.70	25.56	29.37	**29.38**	29.70	**34.81**	28.09
BookCorpus	19.76	19.84	**21.02**	18.11	**18.55**	22.03	20.30	**24.14**
WebText-250k	24.05	**25.25**	20.83	**20.75**	20.51	22.94	24.09	**24.52**
C4-Val	18.96	16.95	**21.69**	**15.89**	14.88	16.40	**21.44**	21.05
SemEval-2018	17.75	**19.05**	0	11.76	**23.91**	**21.05**	10.81	0

joy is found to have high co-occurrence with male. The overall co-occurrence statistics of the gender domain illustrate that negative emotions mostly co-occur with the non-binary gender group, followed by female, and conversely, positive emotions co-occur mostly with the male group. The observations thus clearly dictate imbalanced associations between affective terms and social groups of gender domain, in both pre-training and fine-tuning corpora.

(B) *Emotion Co-occurrence with Racial Domain:* Evaluation results over the racial domain illustrate that the negative emotions *anger* and *sadness* mostly co-occur with African American race group, whereas negative emotion *fear* and the positive emotion *joy* mostly co-occur with European American. But, for all the pre-training corpora, the imbalance of co-occurrence values in the racial domain is comparatively less than the previously discussed gender domain; for example, imbalance in the co-occurrence of all emotions with the racial groups is negligible in the case of WikiEn corpus. Contrary to the observations of pre-training corpora, in fine-tuning corpus SemEval-2018, there exists a large difference in co-occurrence values between African and European American groups. That is, in SemEval-2018, the negative emotions *anger* and *sadness* co-occur with the African American race double the times than European American, indicating highly imbalanced association of *anger* and *sadness* with African American race. Whereas, the co-occurrence of negative emotion *fear* and positive emotion *joy* with European American group is almost thrice African American, again indicating a highly imbalanced association, that of *fear* and *joy* emotions in SemEval-2018 with European American group.

(C) *Emotion Co-occurrence with Religious Domain:* Analysis in the domain of religion shows that *anger* mostly co-occurs with Jewish and *fear* mostly co-occurs with Muslim. Whereas, *joy* is always found to have maximum co-occurrence with Christian. *Sadness* is found to mostly co-occur with Muslim and Jew religious groups than Christian. The results thus shows existence of high co-occurrence between negative emotions *anger*, *fear*, and *sadness* with Muslim and Jew, whereas the positive emotion *joy* with Christian. Moreover, when considering previous observations of gender and racial domains, the imbalance in the religious domain is comparatively higher.

The entire occurrence and co-occurrence analysis over gender, race and religious domains thus consolidate the existence of corpus level affective bias in pre-training and fine-tuning corpora. The extensions of such corpora holding latent affect imbalances, to build computational models may eventually trigger chances of bias in learning models, especially when building large scale contextual pre-trained language models that extract all possible properties of a language.

4. Prediction level affective bias

To identify the existence of prediction level affective bias, if any, in the perspective of large PLMs, we utilize textual emotion detection systems built using popular large PLMs that are fine-tuned using an emotion detection corpus. We evaluate the existence of affective bias in the context of domains gender, race, and religion via different synthetic and non-synthetic paired evaluation sentence corpora and an extensive set of evaluation measures. Details of our investigation, including description and settings of textual emotion detection models based on large PLMs, the method to measure prediction level affective bias with the details of evaluation corpora and measures, and the results and analysis of prediction level affective bias, are given below.

4.1. Textual emotion detection using large PLMs

We formulate the task of textual emotion detection as a four-class classification system with classes being the basic emotions *anger*, *fear*, *joy*, and *sadness*. For this classification task, we utilize pre-trained language models and fine-tune them with an aim to find the best-fit mapping function $f : y = f(x)$ for the fine-tuning data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ with N documents, where x_i indicates ith document in the fine-tuning corpus and y_i indicates the corresponding ground-truth emotion.

The choice of PLMs, GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020), that are utilized in this study to identify affective bias, is motivated by considering their acceptance as relevant and neoteric contextualized models with high performance efficacy towards textual emotion detection (Adoma et al., 2020; Acheampong et al., 2021) and the much related task of sentiment analysis (Zhang et al., 2020; Tabinda Kokab et al., 2022) within the area of affective computing. GPT and BERT are the very popular PLMs that follow the most effective auto-regressive and auto-encoding self-supervised pre-training objectives, respectively, where GPT uses transformer decoder blocks, whereas BERT uses transformer encoder blocks. The autoregressive

Table 5
Fine-tuning corpus statistics.

Emotions	Number of documents	
	Training	Validation
Anger	2089	388
Fear	2641	389
Joy	1906	290
Sadness	1930	397

nature of GPT helps to effectively encode sequential knowledge and achieve good results (Radford et al., 2019). On the other hand, by eliminating the autoregressive objective and alleviating unidirectional constraints through the masked language model pre-training objective, BERT attains powerful bi-directional representations. This ability of BERT to learn context from both sides of a word makes it an empirically powerful state-of-the-art model (Devlin et al., 2019). XLNet brings back the auto-regressive pre-training objective with alternate ways to extract context from both sides of a word and overcome the pretrain-finetune discrepancy of BERT outperforming it in several downstream NLP tasks (Yang et al., 2019). The development of T5 explores the landscape of NLP transfer learning and proposes a unified framework that converts all textual language related problems into the text-to-text format and achieves improved performance (Raffel et al., 2020).

Each pre-trained language model (*PLM*) after fine-tuning and application of *softmax* function at the final layer forms the textual emotion detection model (i.e., *softmax(PLM)*). For each textual document d, the fine-tuned textual emotion detection models predict an emotion class $\hat{e}_{class}$ by finding the highest prediction intensity score $\hat{e}_{score}$ among E classes of emotions (namely *anger*, *fear*, *joy*, and *sadness*, for our task) represented as,

$$\hat{e}_{class}(d) = \underset{k\in 1,2,\ldots,E}{\operatorname{argmax}}\ softmax(PLM(d)) \tag{1}$$

$$\hat{e}_{score}(d) = \max_{k\in 1,2,\ldots,E} softmax(PLM(d)) \tag{2}$$

To fine-tune PLMs and build emotion detection models, we use 24-layered version of the pre-trained BERT, GPT-2, XLNet, and T5 available at HuggingFace,[16] i.e., bert-large-uncased,[17] gpt2-medium,[18] xlnet-large-cased,[19] and t5-large,[20] respectively, and update these architectures by adding a final dense layer of four neurons with softmax activation function on top of the base models to suit our four class classification task. For our study, the choice of GPT-2 instead of the latest version GPT-3 (Brown et al., 2020) is due to its unavailability as an open-source pre-trained model. All four models are fine-tuned using a popular affect detection corpus SemEval-2018 EI-oc (Mohammad et al., 2018) that consists a total of 10 030 data instances for the emotions anger, fear, joy, and sadness. The fine-tuning corpus is split as 8566 data instances for training and 1464 data instances for validation; details of the number of data instances belonging to each emotion category in the train and validation splits are shown in Table 5.

The hyperparameters that can aid the reproducibility of our emotion detection models are, for GPT-2, XLNet, and T5 we use *Adam* optimizer with learning rate 0.000001, categorical crossentropy loss function, and 100 epochs, whereas for BERT the learning rate is 0.00001 and rest of the above mentioned parameters are the same. The batch size is set to 80 for BERT, XLNet, and T5, whereas 64 for GPT-2. The total number of trainable parameters for our BERT, GPT-2, XLNet, and T5 textual emotion detection models come out as 335145988, 354827268, 360272900, and 334943748, respectively. All experiments were conducted on a deep learning workstation equipped with Intel Xeon Silver 4208 CPU at 2.10 GHz, 256 GB RAM, and two GPUs of NVIDIA Quadro RTX 5000 (16 GB for each), using the libraries Tensorflow (version 2.8.0), Keras (version 2.8.0), Transformer (version 4.17.0), and NLTK (version 3.6.5).

[16] https://huggingface.co/.
[17] https://huggingface.co/docs/transformers/model_doc/bert.
[18] https://huggingface.co/docs/transformers/model_doc/gpt2.
[19] https://huggingface.co/docs/transformers/model_doc/xlnet.
[20] https://huggingface.co/docs/transformers/model_doc/t5.

4.2. Measuring prediction level affective bias

The textual emotion detection models, when supplied with a document/sentence, predict as output the emotion class and corresponding emotion intensity of the document/sentence. To identify prediction level affective bias in textual emotion detection models, we input into these models a *sentence pair* that differs only in key terms representing different social groups, with an aim to compare and contrast between emotion predictions of sentences in that pair. For instance, sentence pairs such as '*She made me feel angry*' versus '*He made me feel angry*' that only differ in key terms representing female and male social groups concerning gender domain, or '*African American people can dance very well*' versus '*European American people can dance very well*' that only differ in key terms representing African American and European American social groups concerning racial domain, are input to the models to compare and contrast between emotion predictions of sentences in these pairs. Comparing emotion predictions using such sentence pairs helps to pair-wise analyze and understand whether algorithmic decisions of emotion classification are similar (or different) across different social groups within a domain. Accordingly, to identify prediction level affective bias, we use evaluation corpora that consist of sentence pairs differing only in key terms representing various social groups.

The prediction of emotion class for a sentence is decided by the intensity of emotions predicted by the textual emotion detection model for that sentence. For example, for a prediction $\hat{E}_{score}(d) = \{0.5, 0.2, 0.1, 0.2\}$, the choice of emotion class from the set $E = \{anger, fear, joy, sadness\}$, would be *anger*. Differences in the intensities of emotion predictions between sentences in a pair show existence of affective bias at the intensity level, which when higher enough can alter the prediction of emotion class and thereby cause affective bias at the class level. That is, an unbiased model is expected to predict the same emotion class and intensities for the sentence pairs that only differ in key terms representing different social groups. Hence, to analyze affective bias in the predictions, we utilize class based and intensity based evaluation measures capable of comparing predictions of these sentence pairs. The evaluation corpora and measures are detailed below.

4.2.1. Evaluation corpora

Our choice of bias evaluation corpora is based on the objective to identify affective bias in textual emotion detection models using sentence pairs that only differ in key terms representing social groups, concerning either gender, racial, or religious domain. Suitably, we utilize three different evaluation corpora, Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018), Bias Identification Test in Sentiments (BITS) corpus (Venkit and Wilson, 2021), and Crowdsourced Stereotype Pairs (CSP) corpus (Nangia et al., 2020). Similar to most bias evaluation corpora, EEC and BITS contain template based synthetically created sentences along with ground truth emotions. On the contrary, CSP is a crowd sourced non-synthetic bias evaluation corpus that possesses greater diversity within data in the perspective of context expressed and structure of sentence pairs, but it does not contain ground truth emotions.

EEC consists of a total of 8640 sentences capable of evaluating gender and racial domains, from which we select 8400 sentences for our study after excluding 240 sentences with no emotion words. For the gender domain, the sentence pairs in EEC represent only two social groups, female and male. To tease out affective bias towards the gender domain, we select 1400 sentence pairs. For the racial domain, the sentence pairs represent African American and European American social groups, and we select 2800 sentence pairs to analyze affective bias in racial domain.

BITS consists of a total of 720 sentences capable of evaluating gender and racial domains. In BITS, the sentence pairs for gender domain also include the non-binary social group besides female and male groups. For this study, we select the available 120 sentences for each non-binary, female, and male social group in the gender domain. The racial domain includes social groups such as American Indian, Asian, and Latino, apart from African American and European American groups. For this study, we only select the available small set of 72 sentence pairs representing African American and European American social groups, as we limit our investigation only to these two racial groups.

CSP contains 1508 sentence pairs that cover nine domains, including gender, race, religion, age, nationality, etc., that, unlike EEC and BITS, do not differ 'only' on a single key term. A sample sentence pair in CSP for the racial domain is '*African American people are not very ambitious as a whole*' versus '*European American people are very ambitious as a whole*'. Such sentence pairs violate our evaluation setting of considering sentence pairs that 'only' differ in key terms representing certain social groups, an evaluation strategy to identify bias. But considering the usefulness of such a non-synthetic corpus that suits the real-world context, we manually evaluate all sentence pairs in CSP and modify them (if required) to suit our evaluation settings. An example of modified versions for the above mentioned sentence pair is, '*African American people are not very ambitious as a whole*' versus '*European American people are not very ambitious as a whole*', and '*African American people are very ambitious as a whole*' versus '*European American people are very ambitious as a whole*'. Finally, after such modifications and exclusion of pairs belonging to domains other than gender, race, and religion, we gather 1970 sentences, where the gender domain consists of 263 sentence pairs representing female and male, the racial domain consists of 566 sentence pairs representing African Americans and European Americans, and religious domain consists of 104 sentences each for Christian, Jew, and Muslim social groups.

Even though in some evaluation corpora, certain domains consist of three social groups (e.g. in BITS, the gender domain consists of male, female, and non-binary social groups, in CSP, the religious domain consists of Christian, Jew, and Muslim groups), our evaluation strategies are limited to pair-wise evaluations, to maintain commonality among all the domains. That is, for all the evaluation corpora, from the available set of social groups, we conduct pair-wise evaluations for the pairs, Male versus Female (M × F), Male versus Non-binary (M × Nb), or Female versus Non-binary (F × Nb) in gender domain, European American versus African American (EA × AA) in the racial domain, and Christian versus Muslim (Ch × Mu), Christian versus Jew (Ch × Jw) or Muslim versus Jew (Mu × Jw) in the religious domain.

4.2.2. Evaluation measures

For an evaluation corpus with N sentence pairs, we denote $sp_i^{g_1}$ and $sp_i^{g_2}$ as the ith sentence pair representing two social groups g_1 and g_2 (e.g. Male versus Female), respectively, in a domain (e.g. gender). We evaluate the existence of prediction level affective bias using different measures that rely on class ($\hat{e}_{class}$) and intensity ($\hat{e}_{score}$) predictions of the textual emotion detection models, details follow.

- Demographic Parity (DP): A popular class based measure to quantify group fairness/bias of a classifier system, commonly used to address general affect-agnostic biases like gender bias, racial bias, etc. Du et al. (2021). We utilize this measure to identify the existence of affective bias and check whether the model's emotion classifications are similar (or different) across different social groups within a domain. Accordingly, we say that a textual emotion detection model satisfies demographic parity if,

$$\text{DP} = \frac{P(\hat{e}_{class}(sp^{g_1}) = e | z = g_1)}{P(\hat{e}_{class}(sp^{g_2}) = e | z = g_2)}, \quad e \in E \text{ and } g_1, g_2 \in T \tag{3}$$

  where, $P(\hat{e}_{class}(sp^{g_1}) = e|z = g_1)$ and $P(\hat{e}_{class}(sp^{g_2}) = e|z = g_2)$ indicates the probabilities of the two social groups g_1 and g_2, respectively, to predict an emotion e; g_2 is taken as the group with higher probability (Feldman et al., 2015). E is the set of all emotions, and T is the set of social groups in a domain. Demographic parity advocates the likelihood of emotion prediction outcomes of sentence pairs that differ only in key terms denoting a certain social group should be the same; as a result, DP=1 indicates an ideal unbiased scenario, whereas, lower the values higher the existence of bias. Therefore, we use the general threshold $\tau = 0.80$, lower than which indicates biased predictions (Feldman et al., 2015).
- Average Difference of Prediction Intensity Scores ($avg.\Delta$): An intensity based measure that computes the average difference of emotion prediction intensity scores between the sentence pairs of two social groups in a domain (Kiritchenko and Mohammad, 2018).

$$avg.\Delta = \frac{1}{N}\sum_{i=1}^{N} |\hat{e}_{score}(sp_i^{g_1}) - \hat{e}_{score}(sp_i^{g_2})| \tag{4}$$

  where, $\hat{e}_{score}(sp_i^{g_1})$ and $\hat{e}_{score}(sp_i^{g_2})$ indicates emotion prediction intensity scores corresponding to the social groups g_1 and g_2, respectively, for the ith sentence pair concerning a domain, and N denotes the total number of sentence pairs. That is, $avg.\Delta$ indicates the average dissimilarity in prediction scores between a pair of sentences; 0 indicates perfect similarity, and higher the values more the dissimilarity.
- Prediction Score Significance (p-value): A measure that shows whether dissimilarity in prediction scores between the sentence pairs is statistically significant or not. To compute prediction score significance, we perform a paired statistical significance test, t-Test (Kiritchenko and Mohammad, 2018) over the prediction scores of sentence pairs, $\hat{e}_{score}(sp_i^{g_1})$ and $\hat{e}_{score}(sp_i^{g_2})$, using the conventional significance level, i.e., a p-value of 0.05.
- Average Confidence Score (ACS): A measure that illustrates model bias towards a particular social group using the average ratio between prediction intensity scores of sentence pairs (Nangia et al., 2020), computed as,

$$\text{ACS} = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{\hat{e}_{score}(sp_i^{g_1})}{\hat{e}_{score}(sp_i^{g_2})} \tag{5}$$

  ACS value of an unbiased model will peak around zero, but if it tends to negative values, then the measure indicates that the model prediction intensities of the social group g_1 are higher than g_2, and if it tends to positive values, it indicates that prediction intensities of the social group g_2 are higher than g_1.

4.3. Results and analysis of prediction level affective bias

We examine emotion predictions of each PLM based textual emotion detection system and could observe the existence of affective bias in the predicted emotion classes, as well as their intensities, for gender, race, and religious domains. The sample set of predictions presented in Table 1 is a small subset of these affectively biased emotion predictions from the emotion detection models that employ BERT and T5. More sets of affectively biased predictions from the PLM based textual emotion detection systems, are provided in the supplementary material. In the following subsections, we evaluate the results of each PLM separately.

4.3.1. Affective bias in BERT

Table 6 shows evaluation results observed for the textual emotion detection model built using BERT, analyzing gender, racial and religious domains using three different evaluation corpora EEC, BITS, and CSP, and various evaluation measures. The pairs of social groups addressed by the evaluation corpora within each domain are presented column wise, the measures are presented row wise, and the emotions are grouped across the rows.

Table 6
Results of BERT (Boldface is used to highlight the values of DP < threshold $\tau = 0.80$ and p-values < 0.05).

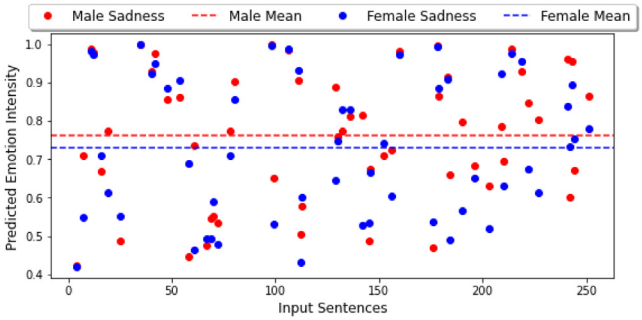
Evaluation	Gender					Race			Religion		
measures	EEC	BITS	CSP	BITS	BITS	EEC	BITS	CSP	CSP	CSP	CSP
	M × F	M × F	M × F	M × Nb	F × Nb	EA × AA	EA × AA	EA × AA	Ch × Mu	Ch × Jw	Mu × Jw
Anger											
DP	0.964	1.000	0.836	0.866	0.867	0.996	0.948	1.000	0.923	0.923	1.000
avg.Δ	0.018	0.016	0.049	0.038	0.030	0.031	0.012	0.052	0.076	0.078	0.100
p-value	**0.003**	**0.036**	**0.037**	**0.047**	0.132	0.417	0.431	0.730	**0.038**	**0.042**	**2e−04**
ACS	0.010	0.017	0.025	0.036	0.020	−0.005	−0.008	−0.001	0.050	−0.084	−0.148
Fear											
DP	0.954	1.000	1.000	0.938	0.938	0.961	1.000	**0.743**	0.857	0.885	0.968
avg.Δ	0.019	0.049	0.086	0.085	0.086	0.049	0.058	0.109	0.076	0.089	0.073
p-value	**9.2e−12**	0.864	0.767	**0.043**	0.063	**5.3e−27**	0.748	**1.2e−6**	**0.044**	0.439	**0.001**
ACS	0.019	−0.010	−0.015	−0.094	−0.088	−0.055	−0.016	−0.123	0.031	−0.041	−0.082
Joy											
DP	0.994	1.000	0.971	1.000	1.000	1.000	1.000	**0.797**	**0.455**	**0.637**	**0.713**
avg.Δ	0.002	9.9e−5	0.072	0.001	0.001	0.005	0.001	0.076	0.148	0.031	0.130
p-value	0.400	0.061	**0.014**	0.360	0.394	**0.002**	0.611	**0.001**	**0.033**	0.425	**0.021**
ACS	−0.001	−5.8e−5	0.064	−0.001	−0.001	−0.004	−1e−4	−0.080	−0.240	−0.022	0.169
Sadness											
DP	0.953	1.000	0.872	0.938	0.938	0.977	0.950	**0.724**	**0.666**	**0.666**	1.000
avg.Δ	0.027	0.013	0.076	0.024	0.033	0.056	0.012	0.116	0.124	0.100	0.051
p-value	**1.8e−4**	**0.045**	**0.019**	0.461	0.156	0.600	0.924	**1e−12**	0.065	0.201	0.146
ACS	−0.020	−0.019	−0.064	0.006	0.022	−0.010	−0.002	0.100	−0.279	−0.169	0.064

(A) *Affective Gender Bias:* Initially, looking into the gender domain, for class based measure DP, throughout all the emotions, we can observe that there is almost no affective bias in the predictions made by BERT between male and female groups when evaluated using the EEC corpus (since, DP > 0.8 in all cases), and ideally no affective bias when evaluated using BITS corpus (since, DP = 1 in all cases). This ideal scenario in BITS might be because BITS is a small corpus containing short-length synthetically created sentences with explicit emotion terms that do not suit the real-world context. When compared to synthetic corpora (EEC and BITS), evaluations using the real-world context and non-synthetic corpus CSP shows more disparity (lower values of DP) between male and female groups for all the emotions except *fear*. For pairs involving non-binary genders, the values of DP are much less than those involving male and female groups of synthetic corpora EEC and BITS, for all emotions except *joy*. This indicate more disparity of male and female groups with non-binary gender, with respect to *anger*, *fear* and *sadness*. Since the evaluation of affective bias in non-binary social groups is only possible with BITS corpus, it may limit the exploration of affective bias towards this group and also the magnitude of affective bias. For the measure DP, when looking across each emotion, the most disparity (lowest value for DP) is observed for *anger* between male versus female when evaluated using CSP corpus, followed by male versus non-binary, and female versus non-binary, for the same emotion, when evaluated using BITS corpus. Whereas, for *joy*, very less disparity is observed across the gender groups. In total, even though disparities are shown by DP, any of the gender pairs do not have values of DP less than the threshold $\tau = 0.80$. Hence DP does not establish the existence of gender affective bias in the predictions of BERT using these evaluation corpora.

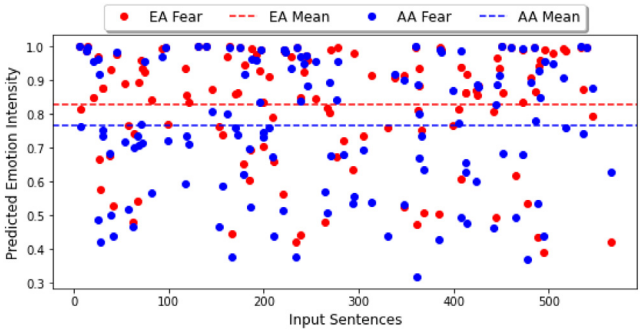
Coming to the intensity based measure avg.Δ in the gender domain, similar to DP, more disparity is observed for male versus female pairs when evaluated using CSP corpus and also for the pairs involving non-binary social groups in BITS, across all the emotions. Different from the measure DP, avg.Δ reports highest disparity for *fear*, but similar to DP, avg.Δ shows very less disparity for *joy*. For the next measure *p*-value, at least one of the evaluation corpora reports values less than 0.05 or statistically significant difference between male and female predictions across the emotions, indicating the existence of affective bias. The *p*-value also shows that difference between male and non-binary predictions for *anger* and *fear* are statistically significant. Analyzing the prediction intensity plots of pairs with statistically significant differences (e.g. Figs. 2(a) and 2(b)), shows that their intensity plots also depict more dispersion between data points as well as more disparity between the corresponding mean values. Conversely, in the plots of sentence pairs with statistically **in**significant differences in prediction intensities (e.g. Fig. 2(c)), there is very less dispersion between data points and less disparity between the mean values. Therefore *p*-value evidently reports the existence of affective bias in emotion prediction intensities of male and female groups with respect to all emotions, and for male and non-binary groups with respect to *anger* and *fear*.

In the case of intensity based measure ACS, for emotion *anger*, the positive values in Male versus Female sentence pairs of EEC, BITS, and CSP indicates that prediction intensities for *anger* are higher for the Female when compared to Male, and positive values in Male versus Non-binary and Female versus Non-binary sentence pairs of BITS indicates that *anger* prediction intensities are higher for the Non-binary group when compared to Male and Female. Similarly, when examining across evaluation corpora, prediction intensities of *fear* and *joy* are higher for Male and Female genders, and prediction intensities of *sadness* are higher for Male and Non-binary genders. Therefore in the gender domain, the measure ACS also indicates affective bias in prediction intensities.

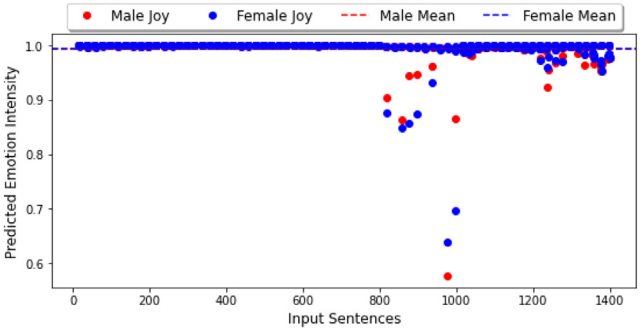
(B) *Affective Racial Bias:* The European and African American racial groups when evaluated using CSP corpus, for the measure DP, shows the presence of affective bias for all emotions except *anger*, where EEC and BITS fail to identify it. Similarly, the avg.Δ disparities among intensity predictions of these racial groups are also much more visible when evaluated using CSP corpus. Either or both, EEC and CSP corpora shows that the difference in intensity predictions of these racial groups are statistically



(a) Plot of *sadness* prediction intensities of M×F in CSP having statistically significant p-value



(b) Plot of *fear* prediction intensities of EA×AA in CSP having statistically significant p-value



(c) Plot of *joy* prediction intensities of M×F in EEC having statistically **in**significant p-value

Fig. 2. Intensity plots of emotion predictions from BERT.

significant with p-values less than 0.05, for all emotions except *anger*, similar to the observations of the measure DP. The measure ACS also shows disparities in prediction intensities between the racial groups, where, for all emotions, prediction intensities of European American race are mostly higher than African American.

(C) *Affective Religious Bias:* In the religious domain, the measure DP evidently shows affective bias in the emotion *joy* with very low values for all three religious pairs and also in *sadness* for Christian versus Muslim and Christian versus Jew pairs. For all the emotions, the values of DP indicate more bias in the Christian versus Muslim and Christian versus Jew sentence pairs than in the Muslim versus Jew pairs. The measure avg.Δ shows that there exist disparities between prediction intensities of religious pairs, and these disparities are found to be comparatively higher than the pairs of gender and racial domains. The p-value indicates statistically significant differences in intensity predictions of *anger* between all three religious pairs. Also, Christian versus Muslim and Muslim versus Jew pairs show statistically significant differences in intensity predictions of all emotions except *sadness*. The measure ACS shows that for BERT *anger* and *fear* prediction intensities are higher for Muslim followed by Christian, and *joy* and *sadness* prediction intensities are higher for Christian followed by Jew.

4.3.2. *Affective bias in GPT-2*

(A) *Affective Gender Bias:* Table 7 shows evaluation results observed for GPT-2 where similar to BERT, no gender affective bias is observed with the measure DP for any of the emotion class predictions. Whereas intensity based disparities are shown by the measure avg.Δ, which is highly visible when evaluated using CSP corpus. The difference in prediction intensities between Male versus Female when evaluated using EEC corpus for all emotions except *joy*, and Male versus Non-binary and Female versus Non-binary when evaluated using BITS corpus for all emotions except *fear*, are statistically significant with p-values < 0.05, indicating the existence of affective bias in emotion prediction intensities. The measure ACS indicates that, in GPT-2, *anger* and *joy* prediction intensities are higher for Male and Female genders, *fear* prediction intensities are higher mainly for Female, and *sadness* prediction intensities are higher mainly for Male gender.

(B) *Affective Racial Bias:* In the racial domain, similar to gender, DP does not show racial affective bias for any of the emotion class predictions, whereas intensity based disparities are shown by the measure avg.Δ. Here also, the disparities for class based measure DP and intensity based measure avg.Δ, are more visible when evaluated using CSP corpus. Whereas BITS reports an ideal unbiased scenario for DP and very low disparity for avg.Δ. The measure p-value reports that the difference in prediction intensities of European and African American races are statistically significant for all emotions except *sadness*. The measure ACS shows that, in GPT-2, prediction intensities of *anger* and *sadness* are mostly higher for African American race, whereas prediction intensities of *fear* and *joy* are mostly higher for European American race.

(C) *Affective Religious Bias:* Unlike gender and race, in the religious domain the class based measure DP reports affective bias (with values of DP < 0.8) in the predictions of all emotions except *fear*. The measure avg.Δ also shows disparities in prediction intensities of religious pairs. The p-values indicate that difference in *fear* prediction intensities for the pairs Christian versus Muslim and Muslim versus Jew are statistically significant. The

Table 7
Results of GPT-2 (Boldface is used to highlight the values of DP < threshold $\tau = 0.80$ and p-values < 0.05).

Evaluation	Gender					Race			Religion		
measures	EEC	BITS	CSP	BITS	BITS	EEC	BITS	CSP	CSP	CSP	CSP
	M × F	M × F	M × F	M × Nb	F × Nb	EA × AA	EA × AA	EA × AA	Ch × Mu	Ch × Jw	Mu × Jw
Anger											
DP	0.992	0.926	0.954	0.960	0.889	0.980	1.000	0.920	**0.600**	0.867	**0.692**
avg.Δ	0.023	0.006	0.039	0.008	0.008	0.038	0.010	0.050	0.059	0.048	0.021
p-value	**2.5e−05**	0.103	0.772	**0.031**	**0.004**	**3.4e−5**	**0.015**	**0.037**	0.580	0.788	0.626
ACS	0.013	0.007	−0.005	−0.006	−0.008	0.011	0.012	0.015	−0.044	−0.018	0.010
Fear											
DP	1.000	1.000	0.991	0.960	0.960	0.996	1.000	0.901	0.883	0.985	0.870
avg.Δ	0.016	0.007	0.058	0.017	0.015	0.030	0.010	0.063	0.139	0.069	0.158
p-value	**0.048**	0.372	0.505	0.917	0.787	**0.012**	0.101	0.183	**6.9e−13**	0.262	**7e−13**
ACS	−0.003	0.002	0.001	3.7e−4	−0.001	−0.011	−0.014	0.005	0.159	−0.040	−0.277
Joy											
DP	0.985	1.000	0.914	1.000	1.000	0.995	1.000	0.936	**0.545**	**0.600**	0.909
avg.Δ	0.008	3.3e−5	0.073	0.001	0.001	0.017	2e−4	0.101	0.114	0.100	0.089
p-value	0.640	0.713	0.761	**0.018**	**0.017**	0.872	0.204	**6.1e−5**	0.110	0.944	0.069
ACS	−7.3e−5	5.3e−6	−0.023	−0.001	−0.001	−0.003	−2e−4	−0.108	0.135	−0.011	−0.129
Sadness											
DP	0.985	0.951	0.927	1.000	0.951	0.996	1.000	0.938	**0.467**	0.933	**0.502**
avg.Δ	0.011	0.002	0.047	0.014	0.014	0.018	0.010	0.055	0.039	0.045	0.045
p-value	**4.5e−29**	0.262	0.313	**0.042**	**0.042**	0.178	0.725	0.283	0.310	0.429	0.343
ACS	−0.012	−0.001	−0.020	−0.013	−0.011	−0.002	0.001	0.006	−0.058	0.028	0.060

measure ACS shows that for GPT-2 *anger* prediction intensities are mostly higher for Christian, *fear* and *joy* prediction intensities are higher for Muslim and Christian, and *sadness* prediction intensities are mostly higher for Jew groups.

4.3.3. Affective bias in XLNet

(A) *Affective Gender Bias:* Table 8 shows evaluation results of XLNet, where the class based measure DP shows negligible affective bias (values of DP is almost one) in emotion predictions of gender pairs, whereas avg.Δ shows disparities in emotion prediction intensities of these pairs. The p-values report that differences between intensity predictions are statistically significant for Male versus Female pairs for all emotions, and also for pairs involving the Non-binary group for emotion *anger*. The measure ACS indicates high *anger* and *fear* prediction intensities for Female and Male genders, and high *joy* and *sadness* prediction intensities for Male and Non-binary genders.

(B) *Affective Racial Bias:* Similar to the gender domain, the measure DP does not confirm class based affective racial bias in XLNet, but avg.Δ shows disparity in intensities of predictions with *p*-value indicating statistically significant differences between prediction intensities of both races, for all emotions. The measure ACS shows that *anger* and *sadness* prediction intensities are higher for African American, whereas *fear* and *joy* prediction intensities are higher for European American race.

(C) *Affective Religious Bias:* In the religious domain, even though the values of DP are less compared to gender and racial domains, it is not sufficient to confirm class based affective religious bias in the emotions except *sadness* whose values are very low and reporting bias. The measure avg.Δ shows disparity in prediction intensities, with *p*-value indicating statistically significant differences between Christian versus Muslim and Muslim versus Jew religious pairs, for *anger* and *sadness*. The measure ACS indicates that *anger* prediction intensities are mostly higher for Muslim religion followed by Christian, *fear* mostly higher for Christian followed by Muslim, and *joy* and *sadness* higher for Christian and Jew.

4.3.4. Affective bias in T5

(A) *Affective Gender Bias:* Table 9 shows evaluation results of T5. In the gender domain, class based measure DP shows affective bias in the predictions of Male versus Female pair for *anger* and *fear* when evaluated using CSP corpus. The avg.Δ measure shows disparities in prediction intensities, and p-values indicate that differences in prediction intensities of Male versus Female pair for all emotions except *fear* and in pairs involving Non-binary gender for emotions *anger* and *fear* are statistically significant. The measure ACS indicates high prediction intensities for *anger*, *joy* and *sadness* mostly by Male gender and high prediction intensities for *fear* mostly by Female and Non-binary genders.

(B) *Affective Racial Bias:* The measure DP does not confirm class based affective racial bias in T5 predictions, whereas avg.Δ shows intensity based affective racial bias, with statistically significant differences in intensity predictions of the racial pairs for all the emotions. ACS indicates prediction intensities of African American race are higher for *anger*, whereas prediction intensities of European American are higher for *fear*, *joy* and *sadness*.

(C) *Affective Religious Bias:* In the religious pairs, the measure DP indicates affective bias in Muslim versus Jew pairs for all emotions, in Muslim versus Christian pairs for all emotions except *anger*, and in Christian versus Jew pairs for *joy*. The avg.Δ shows intensity based disparities in all emotions, and p-values indicate that the differences in prediction intensities are statistically significant in the case of Muslim versus Jew pair for all emotions except *joy* and in Christian versus Jew pair for the emotion *fear*. ACS indicates that *anger* and *joy* prediction intensities are higher for Jew religion followed by Christian, *fear* prediction intensities are higher for Christian followed by Muslim, and *sadness* prediction intensities are higher for Christian followed by Jew.

5. Discussion

5.1. Affective bias - Across the PLMs

This study analyzes affective bias in the predictions of textual emotion detection models at class level and intensity level. In most

Table 8
Results of XLNet (Boldface is used to highlight the values of DP < threshold $\tau = 0.80$ and p-values < 0.05).

Evaluation	Gender					Race			Religion		
measures	EEC	BITS	CSP	BITS	BITS	EEC	BITS	CSP	CSP	CSP	CSP
	M × F	M × F	M × F	M × Nb	F × Nb	EA × AA	EA × AA	EA × AA	Ch × Mu	Ch × Jw	Mu × Jw
Anger											
DP	0.983	1.000	1.000	1.000	1.000	0.976	1.000	0.974	0.825	0.869	0.950
avg.Δ	0.017	0.005	0.053	0.017	0.019	0.048	0.004	0.061	0.115	0.083	0.110
p-value	**1.7e−6**	**0.002**	0.226	**0.035**	**0.014**	**0.041**	0.561	0.063	**0.008**	0.842	**0.001**
ACS	0.015	0.005	−0.028	−0.015	−0.020	−0.021	0.002	0.015	0.077	−0.032	−0.153
Fear											
DP	0.991	1.000	0.989	1.000	1.000	0.988	1.000	0.938	0.810	1.000	0.810
avg.Δ	0.012	0.030	0.080	0.060	0.071	0.038	0.036	0.067	0.054	0.070	0.047
p-value	**0.032**	0.809	0.680	0.667	0.642	0.228	**0.004**	**0.003**	0.561	0.807	0.703
ACS	0.004	−0.001	−0.003	−0.008	−0.013	−0.007	−0.050	−0.062	−0.029	−0.005	−0.019
Joy											
DP	0.993	1.000	0.974	1.000	1.000	0.970	1.000	0.804	0.856	1.000	0.857
avg.Δ	0.010	0.013	0.084	0.006	0.018	0.022	0.009	0.084	0.027	0.077	0.086
p-value	0.457	0.118	**0.028**	0.158	0.125	**0.011**	0.573	**0.024**	0.357	0.410	0.397
ACS	−0.003	−0.018	0.056	0.006	0.019	−0.012	0.004	−0.073	−0.055	0.073	0.133
Sadness											
DP	0.998	1.000	0.989	1.000	1.000	0.997	1.000	0.902	**0.533**	0.833	**0.640**
avg.Δ	0.009	0.003	0.050	0.007	0.008	0.028	0.007	0.083	0.094	0.065	0.104
p-value	**0.013**	**0.010**	0.553	0.203	0.061	0.253	0.075	**5.1e−6**	**0.048**	0.637	**0.010**
ACS	−0.003	−0.003	−0.031	0.002	0.005	−0.004	0.009	0.046	−0.131	0.007	0.124

Table 9
Results of T5 (Boldface is used to highlight the values of DP < threshold $\tau = 0.80$ and p-values < 0.05).

Evaluation	Gender					Race			Religion		
measures	EEC	BIT	CSP	BITS	BITS	EEC	BITS	CSP	CSP	CSP	CSP
	M × F	M × F	M × F	M × Nb	F × Nb	EA × AA	EA × AA	EA × AA	Ch × Mu	Ch × Jw	Mu × Jw
Anger											
DP	0.983	0.966	**0.765**	0.897	0.866	0.933	0.952	0.903	0.968	0.816	**0.790**
avg.Δ	0.039	0.016	0.077	0.021	0.022	0.101	0.004	0.106	0.082	0.113	0.097
p-value	**3.6e−20**	0.530	0.385	**0.017**	**0.043**	**0.001**	0.458	**6.8e−8**	0.118	0.491	**0.041**
ACS	−0.044	0.006	−0.037	−0.029	−0.032	0.005	0.002	0.070	−0.086	0.014	0.064
Fear											
DP	0.994	1.000	**0.778**	0.897	1.000	0.966	1.000	0.867	**0.783**	0.915	**0.717**
avg.Δ	0.017	0.029	0.079	0.079	0.068	0.039	0.067	0.099	0.079	0.148	0.145
p-value	0.309	0.318	0.662	**0.003**	**0.004**	**3.1e−7**	**0.022**	**9.2e−5**	0.602	**0.001**	**2.8e−5**
ACS	0.002	0.008	−0.025	0.071	0.063	−0.035	−0.087	−0.111	−0.005	−0.242	−0.263
Joy											
DP	0.990	1.000	0.848	1.000	1.000	0.961	1.000	0.971	**0.624**	**0.375**	**0.600**
avg.Δ	0.009	2e−4	0.062	1e−4	2.8e−4	0.029	0.009	0.068	0.183	0.001	0.075
p-value	**0.003**	**0.025**	0.885	0.605	0.115	0.122	0.332	**0.001**	0.122	0.468	0.423
ACS	−0.009	−2e−4	−0.025	−1.6e−5	1.8e−4	−0.014	−0.014	−0.078	−0.320	0.001	0.075
Sadness											
DP	0.998	0.973	0.952	0.925	0.900	0.998	0.955	0.972	**0.500**	0.900	**0.450**
avg.Δ	0.023	0.006	0.082	0.009	0.014	0.074	0.007	0.103	0.095	0.118	0.085
p-value	**8.6e−15**	**0.035**	0.689	0.223	0.871	**0.002**	**0.048**	0.957	0.121	0.751	**0.020**
ACS	−0.026	−0.006	−0.027	−0.008	−0.002	−0.040	−0.007	−0.030	−0.150	−0.002	0.099

cases, class based measures that are capable of identifying differences in emotion classes predicted for two different social groups, do not show affective bias, whereas intensity based measures mostly identify the existence of affective bias in predicted emotion intensities. This is because the differences in predicted emotion intensities between the social groups might not be that very high to alter the choice of emotion class predictions, but even then there exists affective bias due to differences in the predicted emotion intensities. When comparing across the PLMs, class based affective gender bias is only observed in T5, whereas intensity based affective gender bias is observed in all the PLMs. Similarly, class based affective racial bias is only observed in BERT, whereas intensity based affective racial bias is observed in all the PLMs. But, in the domain of religion, all four PLMs show high magnitudes of class based and intensity based affective bias, *i.e., compared to gender and race, the religious domain is observed to have high existence of affective bias*. We believe this could be a reflection of comparatively high affect imbalance with respect to the religious domain in the pre-training corpora (from Table 4).

XLNet is observed to have the least class based affective bias, with bias only observed in the case of the religious domain for the emotion *sadness*. XLNet is also observed to have the least intensity based affective bias among all the PLMs when considering the measures avg.Δ (i.e., the top five values of avg.Δ do not have any instance of XLNet) and *p*-value (i.e., the number of instances in XLNet with statistically significant differences are also low). Whereas T5 has the maximum class based biased instances, and also high intensity based affective bias among all the PLMs when considering the measures avg.Δ (i.e., top five values of avg.Δ have three instances of T5) and *p*-value (i.e., the number of instances in T5 with statistically significant differences are also high). BERT also shows class based and intensity based affective bias, nearly similar but comparatively less than T5, followed by GPT-2.

This study explores affective bias in large PLMs that are trained on millions of parameters. However, rapid growth in the data processing technology and plenty availability of data has recently, very quickly evolved the category of large PLMs to being trained on billions of parameters, the PLMs such as LLaMA (Touvron et al., 2023), Flan-T5 XXL (Chung et al., 2022), PaLM (Chowdhery et al., 2023), LaMDA (Thoppilan et al., 2022), etc. All these PLMs have the benefits of huge improvements in their performance, capable of performing several downstream tasks, and supporting multi-lingual and multi-modal data processing. All such large PLMs highly rely on the large availability of massive amounts of textual data especially collected from the web, Wikipedia, Book Corpus, etc. In most cases, which proportion of data is extracted from a source to train a PLM is not fully transparent. For example, LLaMA uses different data proportions from different commonly available data deluges such as CommonCrawl (67.0%), C4 (15.0%), Wikipedia (4.5%), etc., where which data proportions are extracted from each of the sources is not fully transparent. Also, the training data quality is unmanageable and unverifiable by even a large group of human crowd (Navigli et al., 2023), where there are chances of the existence of affective biases in these recent large PLMs. For example, LLaMA is trained on data proportions from several corpora including C4 and Wikipedia, which are already investigated in this study and identified with the affect imbalances. The large PLMs PaLM and LaMDA are also trained on billions of tokens extracted from web pages, books, Wikipedia, and news articles, indicating chances of existence of affective bias. Similarly, Flan-T5 is a variant of the T5, and in this study we observed that T5 has the highest affective bias amongst the PLMs XLNet, BERT, and GPT-2.

5.2. Affect imbalance in corpora and affective bias in predictions

When revisiting the analysis of corpora involved in training PLMs, we have already observed (in Table 4) that these corpora have imbalanced co-occurrences of emotions with certain social groups in gender, racial and religious domains. Further at the prediction level, PLMs that utilize these corpora seems to reflect some of these imbalances hinting at the propagation of affect imbalance in data towards affective bias in predictions. For example, in pre-training and fine-tuning corpora of BERT (i.e., WikiEn, BookCorpus, and SemEval-2018), the emotion *anger* has high co-occurrence with Non-binary and Female groups than Male. This seems to reflect in the predictions of BERT, i.e., the measure ACS shows that prediction intensities of *anger* are higher for Non-binary and Female groups than Male. Some other imbalanced emotion associations that exist in these corpora like *sadness* more associated with Male and Non-binary groups in the gender domain, *joy* more associated with European American racial group, *fear* more associated with Muslim, *joy* more associated with Christian, etc., are also seen to be reflected in the predictions of BERT when evaluated using the measure ACS. Similar to BERT, we can also observe the reflection of corpus level affective bias from pre-training and fine-tuning corpora of GPT-2 (i.e., WebText-250k and SemEval-2018) to the predictions of GPT-2, e.g., (1) high co-occurrence of *fear* with Female and Non-binary genders in the corpora, and high prediction intensities of *fear* for Female and Non-binary genders, (2) high co-occurrence of *anger* with African American race in the corpora, and high prediction intensities of *anger* for African American, (3) high co-occurrence of *fear* with Muslim religion in the corpora, and high prediction intensities of *fear* for Muslim, etc. Such examples of reflection of corpus level affective bias in the predictions of PLMs are also visible in XLNet and T5. These instances give hints that *affect imbalances in the large scale corpora of PLMs may lead to affective bias in the predictions of the models that utilize these PLMs*. Hence, this study further opens the scope for much more nuanced explorations in the direction of affective bias propagation from the corpus to model prediction.

5.3. Societal stereotypes and affective bias

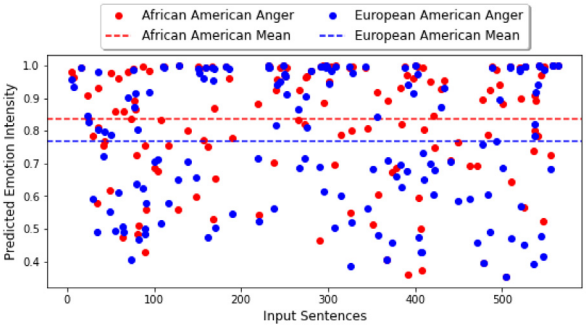
The imbalanced/biased association of emotions with certain social groups within a domain, either at the corpus level or prediction level, reflects several affect-oriented societal stereotypes. Patterns in the training corpora and predictions of PLM based textual emotion detection models showing high association of African American race with *anger* (an example plot of high anger prediction intensities for African American race is presented in Fig. 3(a)) reflect the "Angry Black" stereotype that misrepresents and victimizes blacks as hostile in mainstream American culture and suppress their emotions (Lozada et al., 2022). Another pattern of high association of European American race with *fear* (an example plot of high *fear* prediction intensities for European American is presented in Fig. 3(b)) reflects the existence of stereotypes such as *fear* of crime, residential integration, and racial prejudice among the whites (Skogan, 1995). The high association of Non-binary genders with negative emotions especially *fear*, and very rarely associating with positive emotion *joy*, reflects the societal stigmas like homo-negativity and homophobia against these gender minorities (Hahn et al., 2020). Similarly, the high association of Muslim religion with *fear* (an example plot of high *fear* prediction intensities for Muslim is presented in Fig. 3(c)), which we believe may probably be due to the Islamophobia manifested through text, are inline with the experimental results in Abid et al. (2021b) that reports language generated by GPT-3 (Brown et al., 2020) in the context of the Muslim religion are more associated with violence.

5.4. Effectiveness of evaluation corpora in unveiling affective bias

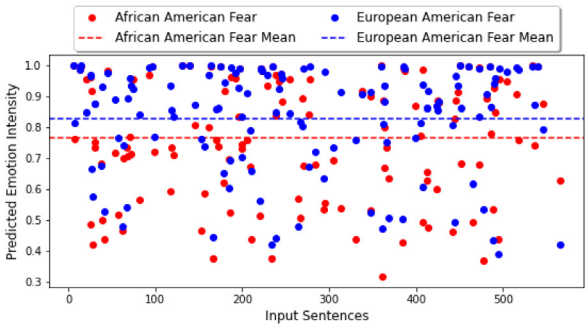
When comparing the capability of the evaluation corpora EEC, BITS, and CSP, we could observe that BITS, with a smaller number of sentence pairs (120 for gender and 72 for race) and explicit emotion terms, is mostly unable to recognize the existence of affective bias in perspective of both class level and intensity level analysis. But even though EEC also has implicit representation of emotion terms similar to BITS, the availability of a large number of sentence pairs (1400 for each domain) eventually helps EEC to identify the existence of affective bias better than BITS. On the other side, even with a smaller number of sentence pairs (263 for gender, 566 for race, 104 for religion), the evaluation corpus CSP helps to identify affective bias to a great extent, and it is the only corpus that unveils class based affective bias in the domains. We believe the non-synthetic and real-world context nature of sentence pairs in CSP could have been advantageous in identifying affective bias. Therefore, upgrading such a corpus with more number of sentence pairs or procuring new evaluation corpora containing non-synthetic real-world sentences, along with corresponding ground truth emotions could eventually help towards comprehensive and rigorous explorations in the direction of identifying affective bias and quantifying its magnitude using ground truth dependent measures like Equal Opportunity (Du et al., 2021).

6. Conclusion

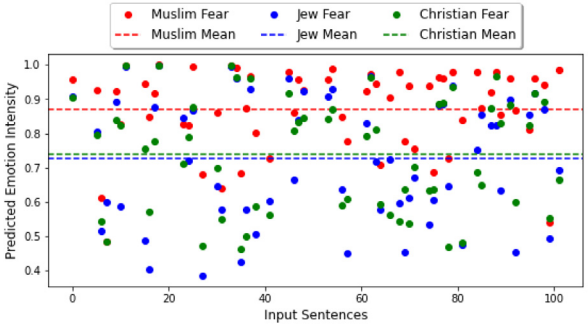
Textual affective analysis and recognition enable efficient ways to encode and understand human emotional states from textual data and yield new opportunities to systems such as business, healthcare, and education by analyzing customers, employees, users, patients, etc., in the context of affective content. Unfair representations of affect in language, i.e. affective bias in such systems discriminate social groups in a domain on the basis of certain emotions while making algorithmic decisions. Affective bias in textual emotion detection systems when deployed in the real world, can harm the ethical trust of these systems and can be potentially threatening to human lives. Hence, analyzing the existence of affective bias in these systems is crucial to avoid huge



(a) High *anger* prediction intensities from T5 for African American race in CSP evaluation corpus reflecting "Angry Black" stereotype



(b) High *fear* prediction intensities from BERT for European American race in CSP evaluation corpus reflecting stereotypes of *fear* in European American



(c) High *fear* prediction intensities from GPT-2 for Muslim religion in CSP evaluation corpus reflecting Islamophobia

Fig. 3. Intensity plots of emotion predictions reflecting societal stereotypes.

disputes and damages in society similar to the adverse effects produced by many other unfair systems such as unfair recidivism prediction.[21]

In this work, we for the first time, to the best of our knowledge, attempted to explore and identify any existence of affective bias in large PLMs, when utilized for the task of textual emotion detection, with respect to the domains gender, race, and religion. For the study, we used BERT, GPT-2, XLNet, and T5 considering their popularity and wide applicability in textual emotion detection and many other related tasks. As algorithmic bias has its roots from data bias, we started our exploration of affective bias by analyzing the imbalanced distribution of affect in the pre-training corpora of these PLMs i.e., WikiEn, BookCorpus, WebText-250, and C4-Val, and SemEval-2018 used to fine-tune the emotion detection models. Later, we analyzed the existence of affective bias in the predictions of fine-tuned emotion detection models built using these large PLMs. Evaluations are performed to analyze affective bias in the predicted emotion classes and corresponding intensities of social groups within a domain using three different evaluation corpora and various class based and intensity based evaluation measures. Our wide set of experiments and evaluation strategies confirm the existence of affect imbalance in large scale corpora and affective bias in emotion predictions of the PLMs, with affective bias mostly higher for T5 compared to the other PLMs. The high association of emotion *anger* with African American race, *joy* with European American race, *fear* with the Muslim religion, etc., are some examples of affective bias. Religious domain reports more biased instances, compared to gender and race, for all the PLMs. Our results also demonstrated that the biased predictions of the models are inclined with patterns of affect imbalance in the corpora, and both these reflect certain affect-oriented societal stereotypes, hinting at the propagation of affective bias towards predictions of the PLMs. To aid future research, we shall make publicly available all the relevant materials including the pre-processed pre-training and fine-tuning corpora, evaluation corpora modified to suit our task, list of affective terms and target terms for corpus level analysis, source code, and fine-tuned textual emotion detection models along with their emotion class and intensity predictions, at https://github.com/anoopkdcs/affective_bias_in_plm and https://dcs.uoc.ac.in/cida/projects/ac/affective-bias.html along with the publication.

6.1. Future work

The proposed study explores corpus level affective bias using a simple approach to analyzing the distributions of affective target terms in the corpora. In the future, we are planning to conduct a more nuanced exploration towards the corpus level affective bias in the context of various facets such as the time of creation of corpora, people behind corpora, languages and cultures (Navigli et al., 2023). Recent affect agnostic bias analysis studies explore bias in the context of causality (Su et al., 2022); therefore to further explore the relationship between the corpus characteristics and model bias we are also planning to conduct causality based affective bias analysis.

In context the model predictions, the observations of affective bias and its magnitudes in this study are dependent on the choice of evaluation corpora and measures, i.e., certain instances of 'no affective bias' or marginal magnitudes of affective bias may also be due to the limited capability of evaluation corpora and measures to unveil the actual latent affective bias that exists in the model. Therefore in the future, we are considering extending the study with a set of real-world

[21] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=nD-X136_tDm0nh1l4Xtv0LbpjY_BSO3u.

context evaluation corpora, for example, by expanding CSP in terms of the number of sentences and also by procuring ground truth emotions that allow applying other evaluation measures like Equal Opportunity (Du et al., 2021). Beyond analyzing each sentence pair in a domain separately, we are looking into the ways to simultaneously analyze sentences representing various social groups in a domain, for example, analyzing sentence triplets like Male versus Female versus Non-binary.

A very recent and relevant work that addresses a similar line of thoughts in the context of affect agnostic bias in PLMs from pre-training data to language models to downstream tasks in the political domain is explained in Feng et al. (2023). In the backdrop of this work, we observe a wide scope of exploring *political affective bias* in large PLMs. Because, there are works in the literature that give hints that emotions such as anger, disgust, or fear are more frequent in the predictions of republicans' (right-leaning) posts, whereas love or sadness are more often predicted for democrats' (left-leaning) posts (Huguet Cabot et al., 2020).

Our initial attempt to identify affective bias in textual emotion detection models that utilize large PLMs, opens up the vast future scope towards identifying affective bias in the other very recent large PLMs such as LLAMA, Flan-T5 XXL, PaLM, LaMDA, etc. There also exists a wide scope for affective bias mitigation, which we believe, can be better achieved by adopting more convenient solutions that utilize constraints while fine-tuning the prediction system (i.e., in-processing) and post-processing, rather than retraining or fine-tuning the PLM based affect prediction systems with unbiased corpora which are expensive and cumbersome (Hooker, 2021).

CRediT authorship contribution statement

Anoop Kadan: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Deepak P.:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – review & editing, Validation. **Sahely Bhadra:** Formal analysis, Methodology, Supervision, Writing – review & editing, Validation. **Manjary P. Gangan:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Lajish V.L.:** Supervision, Writing – review & editing.

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.

Acknowledgments

The authors would like to thank the authors of Tan and Celis (2019) for making their source codes publicly available and the authors of Kiritchenko and Mohammad (2018), Venkit and Wilson (2021) and Nangia et al. (2020) for making their evaluation corpora publicly available. The authors would like to thank Chanjal V.V., Master's student (2018–20) of the Department of Women Studies, University of Calicut for her involvement and cooperation to create the list of target terms related to non-binary gender to conduct the corpus level experiments. The first author would like to thank Indian Institute of Technology Palakkad for organizing the GIAN course on Fairness in Machine Learning. The third author would like to thank the Department of Science and Technology (DST) of the Government of India for financial support through the Women Scientist Scheme-A (WOS-A) for Research in Basic/Applied Science under the Grant SR/WOS-A/PM-62/2018.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.nlp.2024.100062.

References

Abid, A., Farooqi, M., Zou, J., 2021a. Large language models associate muslims with violence. Nat. Mach. Intell. 3 (6), 461–463. http://dx.doi.org/10.1038/s42256-021-00359-2.

Abid, A., Farooqi, M., Zou, J., 2021b. Persistent anti-muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, pp. 298–306, URL: https://doi.org/10.1145/3461702.3462624.

Acheampong, F.A., Nunoo-Mensah, H., Chen, W., 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif. Intell. Rev. 54 (8), 5789–5829. http://dx.doi.org/10.1007/s10462-021-09958-2.

Adoma, A.F., Henry, N.-M., Chen, W., 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing. ICCWAMTIP, pp. 117–121. http://dx.doi.org/10.1109/ICCWAMTIP51612.2020.9317379.

Anoop, K., Gangan, M.P., Deepak, P., Lajish, V.L., 2022. Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias. In: Responsible Data Science. Springer Nature, Singapore, pp. 13–45. http://dx.doi.org/10.1007/978-981-19-4453-6_2.

Ashley, W., 2014. The angry black woman: The impact of pejorative stereotypes on psychotherapy with black women. Soc. Work Public Health 29 (1), 27–34. http://dx.doi.org/10.1080/19371918.2011.619449.

Bhaskaran, J., Bhallamudi, I., 2019. Good secretaries, bad truck drivers? Occupational gender stereotypes in sentiment analysis. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Italy, pp. 62–68. http://dx.doi.org/10.18653/v1/W19-3809.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., Kalai, A., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS '16, Curran Associates Inc., Red Hook, NY, USA, pp. 4356–4364, URL: https://dl.acm.org/doi/10.5555/3157382.3157584.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901, URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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

Center, T.S., 2022. LGBTQIA+ terminology. URL: https://www.umass.edu/stonewall/sites/default/files/documents/allyship_term_handout.pdf. Accessed: 4-7-2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2023. Palm: Scaling language modeling with pathways. J. Mach. Learn. Res. 24 (240), 1–113, URL: https://www.jmlr.org/papers/volume24/22-1144/22-1144.pdf.

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al., 2022. Scaling instruction-finetuned language models. http://dx.doi.org/10.48550/arXiv.2210.11416, arXiv preprint arXiv:2210.11416.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 2017. Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, Association for Computing Machinery, New York, NY, USA, pp. 797–806. http://dx.doi.org/10.1145/3097983.3098095.

Dale, R., 2019. Law and word order: NLP in legal tech. Nat. Lang. Eng. 25 (1), 211–217. http://dx.doi.org/10.1017/S1351324918000475.

De Choudhury, M., Counts, S., Gamon, M., 2012. Not all moods are created equal! exploring human emotional states in social media. In: Proceedings of the International AAAI Conference on Web and Social Media. Vol. 6, pp. 66–73, URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14279.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

Díaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D., 2018. Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–14, URL: https://doi.org/10.1145/3173574.3173986.

Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L., 2018. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, Association for Computing Machinery, New York, NY, USA, pp. 67–73. http://dx.doi.org/10.1145/3278721.3278729.

Du, M., Yang, F., Zou, N., Hu, X., 2021. Fairness in deep learning: A computational perspective. IEEE Intell. Syst. 36 (4), 25–34. http://dx.doi.org/10.1109/MIS.2020.3000681.

Eagly, A.H., Steffen, V.J., 1984. Gender stereotypes stem from the distribution of women and men into social roles. J. Pers. Soc. Psychol. 46 (4), 735, doi:https://psycnet.apa.org/doi/10.1037/0022-3514.46.4.735.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S., 2015. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, Association for Computing Machinery, New York, NY, USA, pp. 259–268. http://dx.doi.org/10.1145/2783258.2783311.

Feng, S., Park, C.Y., Liu, Y., Tsvetkov, Y., 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp. 11737–11762. http://dx.doi.org/10.18653/v1/2023.acl-long.656, URL: https://aclanthology.org/2023.acl-long.656.

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J., 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Natl. Acad. Sci. 115 (16), E3635–E3644.

Guo, W., Caliskan, A., 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, pp. 122–133, URL: https://doi.org/10.1145/3461702.3462536.

Hahn, H., Seager van Dyk, I., Ahn, W.-Y., 2020. Attitudes toward gay men and lesbian women moderate heterosexual adults' subjective stress response to witnessing homonegativity. Front. Psychol. 10, 2948. http://dx.doi.org/10.3389/fpsyg.2019.02948.

He, P., Liu, X., Gao, J., Chen, W., 2021. DEBERTA: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=XPZIaotutsD.

Hooker, S., 2021. Moving beyond "algorithmic bias is a data problem". Patterns 2 (4), 100241. http://dx.doi.org/10.1016/j.patter.2021.100241, URL: https://www.sciencedirect.com/science/article/pii/S2666389921000611.

Hovy, D., Prabhumoye, S., 2021. Five sources of bias in natural language processing. Lang. Linguist. Compass 15 (8), e12432. http://dx.doi.org/10.1111/lnc3.12432, URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., Kohli, P., 2020. Reducing sentiment bias in language models via counterfactual evaluation. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp. 65–83. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.7.

Huguet Cabot, P.-L., Dankers, V., Abadi, D., Fischer, A., Shutova, E., 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In: Cohn, T., He, Y., Liu, Y. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp. 4479–4488. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.402, URL: https://aclanthology.org/2020.findings-emnlp.402.

Kaneko, M., Bollegala, D., 2022. Unmasking the mask – evaluating social biases in masked language models. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, BC, Canada, http://dx.doi.org/10.1609/aaai.v36i11.21453.

Kiritchenko, S., Mohammad, S., 2018. 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.

Liang, P.P., Wu, C., Morency, L.-P., Salakhutdinov, R., 2021. Towards understanding and mitigating social biases in language models. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 6565–6576, URL: https://proceedings.mlr.press/v139/liang21a.html.

Lozada, F.T., Riley, T.N., Catherine, E., Brown, D.W., 2022. Black emotions matter: Understanding the impact of racial oppression on black youth's emotional development: Dismantling systems of racism and oppression during adolescence. J. Res. Adolesc. 32 (1), 13–33. http://dx.doi.org/10.1111/jora.12699.

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 International Publishing, Cham, pp. 189–202. http://dx.doi.org/10.1007/978-3-030-62077-6_14.

Mao, R., Liu, Q., He, K., Li, W., Cambria, E., 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. IEEE Trans. Affect. Comput. 1–11. http://dx.doi.org/10.1109/TAFFC.2022.3204972.

May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R., 2019. On measuring social biases in sentence encoders. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 622–628. http://dx.doi.org/10.18653/v1/N19-1063.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T., Trajanov, D., 2020. Evaluation of sentiment analysis in finance: From lexicons to transformers. IEEE Access 8, 131662–131682. http://dx.doi.org/10.1109/ACCESS.2020.3009626.

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S., 2018. SemEval-2018 task 1: Affect in tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana, pp. 1–17. http://dx.doi.org/10.18653/v1/S18-1001, URL: https://aclanthology.org/S18-1001.

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

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

Navigli, R., Conia, S., Ross, B., 2023. Biases in large language models: Origins, inventory, and discussion. J. Data Inf. Qual. 15 (2), http://dx.doi.org/10.1145/3597307.

Plant, E.A., Hyde, J.S., Keltner, D., Devine, P.G., 2000. The gender stereotyping of emotions. Psychol. Women Q. 24 (1), 81–92. http://dx.doi.org/10.1111/j.1471-6402.2000.tb01024.x.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9, URL: https://openai.com/blog/better-language-models/.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67, URL: http://jmlr.org/papers/v21/20-074.html.

Rahman, M.M., Siddiqui, F.H., 2019. An optimized abstractive text summarization model using peephole convolutional LSTM. Symmetry 11 (10), http://dx.doi.org/10.3390/sym11101290, URL: https://www.mdpi.com/2073-8994/11/10/1290.

Rahman, M.M., Siddiqui, F.H., 2021. Multi-layered attentional peephole convolutional LSTM for abstractive text summarization. ETRI J. 43 (2), 288–298. http://dx.doi.org/10.4218/etrij.2019-0016, URL: https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.2019-0016.

Raza, S., Garg, M., Reji, D.J., Bashir, S.R., Ding, C., 2024. Nbias: A natural language processing framework for BIAS identification in text. Expert Syst. Appl. 237, 121542. http://dx.doi.org/10.1016/j.eswa.2023.121542, URL: https://www.sciencedirect.com/science/article/pii/S0957417423020444.

Rozado, D., 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. PLoS One 15 (4), 1–26. http://dx.doi.org/10.1371/journal.pone.0231189.

Shen, J.H., Fratamico, L., Rahwan, I., Rush, A.M., 2018. Darling or babygirl? investigating stylistic bias in sentiment analysis. Proc. FATML URL: https://www.fatml.org/media/documents/darling_or_babygirl_stylistic_bias.pdf.

Shields, S.A., 2002. Speaking from the Heart: Gender and the Social Meaning of Emotion. Cambridge University Press.

Skogan, W.G., 1995. Crime and the racial fears of white Americans. Ann. Am. Acad. Political Soc. Sci. 539 (1), 59–71. http://dx.doi.org/10.1177/0002716295539001005.

Soni, S., Roberts, K., 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 5532–5538, URL: https://aclanthology.org/2020.lrec-1.679.

Staiano, J., Guerini, M., 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 427–433. http://dx.doi.org/10.3115/v1/P14-2070, URL: https://aclanthology.org/P14-2070.

Su, C., Yu, G., Wang, J., Yan, Z., Cui, L., 2022. A review of causality-based fairness machine learning. Intell. Robot. 244–274. http://dx.doi.org/10.20517/ir.2022.17.

Subramanian, S., Rahimi, A., Baldwin, T., Cohn, T., Frermann, L., 2021. Fairness-aware class imbalanced learning. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 2045–2051. http://dx.doi.org/10.18653/v1/2021.emnlp-main.155, URL: https://aclanthology.org/2021.emnlp-main.155.

Suresh, H., Guttag, J., 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3465416.3483305.

Sweeney, C., Najafian, M., 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In: Proceedings of the

2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 359–368. http://dx.doi.org/10.1145/3351095.3372837.

Tabinda Kokab, S., Asghar, S., Naz, S., 2022. Transformer-based deep learning models for the sentiment analysis of social media data. Array 14, 100157. http://dx.doi.org/10.1016/j.array.2022.100157, URL: https://www.sciencedirect.com/science/article/pii/S2590005622000224.

Tan, Y.C., Celis, L.E., 2019. Assessing social and intersectional biases in contextualized word representations. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp. 13230–13241, URL: https://dl.acm.org/doi/10.5555/3454287.3455472.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al., 2022. Lamda: Language models for dialog applications. http://dx.doi.org/10.48550/arXiv.2201.08239, arXiv preprint arXiv:2201.08239.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. URL: https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/.

Trinh, T.H., Le, Q.V., 2018. A simple method for commonsense reasoning. http://dx.doi.org/10.48550/arXiv.1806.02847, arXiv preprint arXiv:1806.02847.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A.D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., Dutta, R., 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. J. Biomed. Inform. 88, 11–19. http://dx.doi.org/10.1016/j.jbi.2018.10.005, URL: https://www.sciencedirect.com/science/article/pii/S1532046418302016.

Venkit, P.N., Wilson, S., 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. http://dx.doi.org/10.48550/arXiv.2111.13259, arXiv preprint arXiv:2111.13259.

Vittengl, J.R., Holt, C.S., 1998. A time-series diary study of mood and social interaction. Motiv. Emot. 22 (3), 255–275. http://dx.doi.org/10.1023/A:1022388123550.

Waterloo, S.F., Baumgartner, S.E., Peter, J., Valkenburg, P.M., 2018. Norms of online expressions of emotion: Comparing facebook, Twitter, instagram, and WhatsApp. New Media Soc. 20 (5), 1813–1831. http://dx.doi.org/10.1177/1461444817707349, PMID: 30581358.

Yang, Z., Asyrofi, M.H., Lo, D., 2021. BiasRV: Uncovering biased sentiment predictions at runtime. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 1540–1544. http://dx.doi.org/10.1145/3468264.3473117.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V., 2019. XLNet: Generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp. 5753–5763, URL: https://dl.acm.org/doi/10.5555/3454287.3454804.

Zhang, L., Fan, H., Peng, C., Rao, G., Cong, Q., 2020. Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. Healthcare 8 (3), http://dx.doi.org/10.3390/healthcare8030307, URL: https://www.mdpi.com/2227-9032/8/3/307.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.-W., 2019. Gender bias in contextualized word embeddings. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 629–634. http://dx.doi.org/10.18653/v1/N19-1064, URL: https://aclanthology.org/N19-1064.

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

Zhiltsova, A., Caton, S., Mulway, C., 2019. Mitigation of unintended biases against non-native english texts in sentiment analysis. In: Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, December 5-6, 2019. In: CEUR Workshop Proceedings, vol. 2563, CEUR-WS.org, pp. 317–328, URL: http://ceur-ws.org/Vol-2563/aics_30.pdf.

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