IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 December 2023, accepted 26 January 2024, date of publication 30 January 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360306

SURVEY

Shortcut Learning Explanations for Deep Natural Language Processing: A Survey on Dataset Biases

VARUN DOGRA[1], SAHIL VERMA[2], (Senior Member, IEEE), KAVITA[2], (Senior Member, IEEE), MARCIN WOŹNIAK[3], JANA SHAFI[4], AND MUHAMMAD FAZAL IJAZ[5]

[1]School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab 144411, India
[2]Uttaranchal University, Dehradun 248007, India
[3]Faculty of Applied Mathematics, Silesian University of Technology, 44100 Gliwice, Poland
[4]Department of Computer Engineering and Information, College of Engineering in Wadi Alddawasir, Prince Sattam Bin Abdulaziz University, Wadi Alddawasir 11991, Saudi Arabia
[5]School of IT and Engineering, Melbourne Institute of Technology, Melbourne, VIC 3000, Australia

Corresponding authors: Marcin Woźniak (marcin.wozniak@polsl.pl) and Muhammad Fazal Ijaz (mfazal@mit.edu.au)

This work was supported in part by the Silesian University of Technology under Grant 09/010/RGJ24/0031, and in part by Prince Sattam bin Abdulaziz University under Project PSAU/2023/R/1445.

ABSTRACT The introduction of pre-trained large language models (LLMs) has transformed NLP by fine-tuning task-specific datasets, enabling notable advancements in news classification, language translation, and sentiment analysis. This has revolutionized the field, driving remarkable breakthroughs and progress. However, the growing recognition of bias in textual data has emerged as a critical focus in the NLP community, revealing the inherent limitations of models trained on specific datasets. LLMs exploit these dataset biases and artifacts as expedient shortcuts for prediction. The reliance of LLMs on dataset bias and artifacts as shortcuts for prediction has hindered their generalizability and adversarial robustness. Addressing this issue is crucial to enhance the reliability and resilience of LLMs in various contexts. This survey provides a comprehensive overview of the rapidly growing body of research on shortcut learning in language models, classifying the research into four main areas: the factors of shortcut learning, the origin of bias, the detection methods of dataset biases, and understanding mitigation strategies to address data biases. The goal of this study is to offer a contextualized, in-depth look at the state of learning models, highlighting the major areas of attention and suggesting possible directions for further research.

INDEX TERMS Dataset biases, deep learning, natural language processing, shortcut learning, transfer learning.

I. INTRODUCTION

In the field of Natural Language Processing, pre-trained large language models have gained significant attention. Notably, language models like BERT [1], Roberta [2], and GPT-3 [3] have shown their abilities in performing several high-level NLP tasks, such as natural language inference, question answering, text summarization, sentiment analysis, etc. BERT and its variants are trained using masked language modeling objectives, where a portion of the input text is randomly masked, and the model is tasked with predicting the original masked tokens (e.g. "I love to [MASK] on the beach"). The model leverages contextual information from both the left and right sides to understand the masked token. Formally, given an input sequence as $w = [w_1, w_2, \ldots, w_n]$ and a position $1 \leq i \leq n$, the model estimates the token by considering its neighboring left and right contexts by $p(w) = p(w_i \mid w_1, w_2, \ldots, w_{i-1}, w_{i+1}, w_{i+2}, \ldots, w_n)$.

While this training objective is effective in capturing contextual relationships and improving language understanding, it introduces a challenge known as *shortcut learning*. In the case of masked language models, the models may learn to

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

rely on prominent patterns or contextual clues rather than deeply understanding the semantics or meaning of the text. For example, the models could learn associations between specific words or phrases with certain labels or predictions, without truly grasping the underlying concepts.

Secondly, LLMs often rely on transfer learning, where they are pre-trained on a large corpus of text data and fine-tuned on specific downstream tasks. During this fine-tuning process, the model adapts its parameters to the new task while leveraging the knowledge gained from the pre-training phase. The adaptation of large language models to downstream tasks typically involves employing three conventional approaches as shown in Figure 1. The general principle states that increasing the number of fine-tuned layers typically leads to improved performance. So, the approach of employing pre-trained LLMs to downstream tasks as shown in Fig. 1(iii) may produce higher accuracy than the other approaches shown in Fig. 1(i) and 1(ii). However, if the fine-tuning process is not carefully managed, shortcut learning can occur.

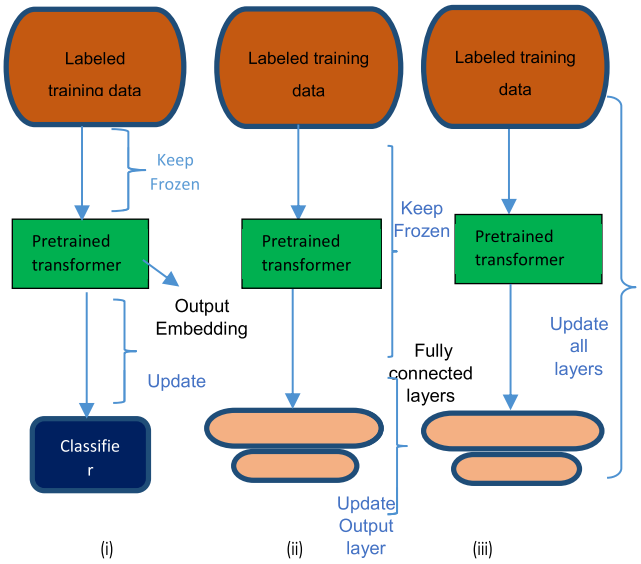

FIGURE 1. (i) represents a feature-based approach for adapting the model on downstream tasks, (ii) and (iii) represent fine-tuning approaches.

Thirdly, LLMs have exhibited exceptional proficiency in addressing significant challenges and achieving outstanding performance across a wide range of Natural Language Processing (NLP) problems. However, despite their remarkable performance on in-distribution test data, face challenges in terms of generalization when applied to out-of-distribution (OOD) test data. These models exhibit reduced performance on OOD samples and are susceptible to various adversarial attacks, which collectively contribute to their vulnerability and low robustness [4]. The significant contributing factor to the low robustness observed in LLMs is Shortcut Learning.

In all these cases, the learning of shortcuts in natural language understanding (NLU) models may be influenced by multiple factors including the dataset biases, training and fine-tuning processes, as well as potential biases present in the embedding spaces [5]. A significant area of concern regarding large language models is the possibility of biases and unfairness in their predictions, particularly when trained on datasets that themselves contain biases [6]. If the training data is skewed or biased towards certain patterns, topics, or demographics, LLMs may inadvertently learn and amplify these biases. These shortcuts can lead to models relying on biases in the data rather than developing a deep understanding of the underlying semantics. As a result, the models tend to exhibit biased behavior and may make unfair or inaccurate predictions when applied to diverse or underrepresented samples. This raises the following research questions for us to consider.

1. Can transfer learning approaches be used to train pre-trained LLMs that are less prone to shortcut learning by actively encouraging the model to acquire task-specific features rather than relying on shortcuts?
2. Can techniques be developed to prevent pre-trained LLMs from being overly reliant on certain features that they frequently utilize as shortcuts?
3. How can regularization approaches be improved to reduce shortcut learning in pre-trained LLMs without compromising their capacity to recognize complex linguistic patterns?
4. Are there pre-training process changes that can be made to reduce or avoid shortcut learning in pre-trained LLMs?

Some studies have aimed to propose modifications to the training process that effectively address the issue, by introducing task-specific regularization techniques [7], [8], adversarial examples [9], or applying data augmentation techniques during training [10], ultimately promoting a more precise and unbiased understanding of natural language. To ensure that algorithm-guided outcomes are fair and unbiased, three formal definitions of fairness have emerged from the literature: (1) anti-classification, which states that protected attributes such as race, gender, and their proxies are not explicitly used to make decisions; (2) classification parity, which states that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; and (3) calibration, which states that outcomes are independent of risk estimates and unfairness in such models' predictions, particularly when trained on biased datasets.

Stronger anti-classification theories have been put out to prevent the use of unprotected features as substitutes for protected attributes [11]. These more robust ideas seek to address the problem of bias and discrimination in machine learning systems. In the subsequent section of the paper, we explore specific studies focusing on dataset biases in machine learning systems. These studies mention the motivations behind identifying and understanding how machine learning system behaviors (particularly language models for solving NLP problems) can be perceived as harmful.

The robustness of the models has been significantly harmed by shortcut learning due to dataset biases, drawing

increased attention from the NLP community to find a solution. The background of the Dataset biases in the context of shorting learning problems in LLMs is given in Section II. The existing studies focus on Shortcut learning explanations are reviewed in Section III, dataset-bias detection methods in Section IV, mitigation methods in Section V, and a comparison of bias measures in Section VI.

II. BACKGROUND: ORIGIN OF DATASET BIASES

Recent studies have brought to light significant biases in Natural Language Processing (NLP), showcasing potential harm. However, a common shortcoming in many of these studies is a lack of critical engagement with the fundamental definition and understanding of the ''bias'' [12]. The term is frequently employed broadly, encompassing various system behaviors in NLP, such as gender and racial bias. Even when scrutinizing bias in NLP systems designed for a common objective, diverse research endeavors may exhibit distinct understandings and conceptualizations of what constitutes bias. For instance, the behaviors of systems, such as deeming the statement ''You are a good woman'' as sexist when trained on a particular dataset [13]. Additionally, reliance on the Equity Evaluation Corpus has uncovered instances where certain methods consistently predict sentiment intensity levels slightly higher for specific races or genders [14]. This diversity in biases underscores the complexity of addressing bias in NLP systems and highlights the necessity for nuanced considerations in research and mitigation efforts.

From the examples, it is clear that the performance and behavior of NLP models are significantly influenced by the datasets that are used, as both researchers and practitioners have come to understand over time [15]. The ability of the model to generalize, identify patterns, and make reliable predictions can be significantly impacted by the quality, size, diversity, and representativeness of the data. The model may pick up on and reinforce biases present in the training data, producing biased results or reaffirming preexisting societal biases. The history of dataset development in the fields of data science and machine learning can be seen as a depiction of resistance against perceived unfairness and bias [16].

Each successive dataset has developed in response to the biases and constraints present in preceding datasets, to highlight the context more thoroughly and objectively. On the one hand, this change in the way datasets are being developed may be a sign of advancement. On the other side, a slight vicious cycle was also apparent. We as a community consistently reject the present datasets because we believe they are biased. Yet each time we construct a new dataset, it turns out to be biased in the same way, although in a little different way. We are destined to make the same mistakes over and over again, therefore what seems to be lacking is a clear knowledge of the different forms and origins of bias [17].

Biases are divided into two types; dataset bias and modeling bias. CNNs trained on ImageNet typically identify images based on texture rather than shape is an example of modeling bias [18]. It is also found that CNNs learn to classify by shape at least as quickly as by texture when trained on datasets containing images having conflicting shapes and textures. On the other hand, unbalanced sample sizes for each category, correlations between categories and unrelated attributes, and data distribution that reflects social assumptions are a few examples of dataset biases.

Cognitive sciences have been researching various biases for many years. Bias has been recognized for a long time as a natural human directorial approach. Consider the possibility that inferential judgments guide intuitive predictions. By using this criterion, individuals make predictions based on what the evidence suggests will happen. Therefore, contrary to the logic of statistical prediction, intuitive forecasts are immune to the validity of the evidence or the prior probability of the result. An imperfect induction method or learning from others are two ways that bias can be developed. A bias may lead to a manner of thinking that departs from actual logic in any event [19].

In a machine learning environment, when a model's results do not discriminate based on specific features, it is said to be unbiased. The confusion matrices for various targeted classes can be used to estimate bias [16]. In other words, we can calculate confusion matrices and derived rates for each subset of data that was created by segmenting the complete collection of samples on a particular feature. If these rates are significantly different from one another, this may indicate the lack of unbiased behavior on the part of the prediction system, or rather, an obvious bias in decision-making based on the importance of that particular attribute. Divergences in prediction rates among various population groups have been studied using a variety of measurements, and it is now obvious how to interpret these measures in light of each system's purpose. Machine learning algorithms for prediction have been applied in important decisions affecting human life for many years. The predictions made by the algorithms utilizing them would also differ greatly since certain formalizations of fairness can contradict others. Therefore, from a practical point of view, measures should have been studied on how fairness is formalized in the literature of machine learning and the effects of various formalizations [20].

The need to provide a wider context and a more realistic depiction of real-world settings where machine learning models are deployed has motivated the creation of new datasets. Dataset building has been crucial in reducing biases and progressing the industry towards more equitable and reliable machine learning systems by aiming for fairness, inclusion, and a more thorough representation of many groups and perspectives. It is important to keep in mind that dataset development is still a dynamic, iterative process. By continually updating and enhancing databases, researchers aim to mitigate biases and produce more impartial and reliable outcomes in various applications of machine learning. The following section delves into a more detailed survey of the recent insights into the factors of shortcut learning in LLMs.

III. RELATED WORKS

In recent studies, the limitations and drawbacks of large language models have gained attention. Notably, when confronted with shifts in input distribution, such as those encountered during domain adaptation, with out-of-distribution examples, or when faced with novel scenarios, LLMs exhibit challenges in adapting and providing reliable predictions. This suggests that LLMs may encounter difficulties in adjusting to new input distributions, potentially relying on biases ingrained during training [21], [22], [23]. One type of bias observed is the strong co-occurrence correlation between certain class labels and specific lexical features. These features typically include low-level functional words like stop words, numbers, negation words, and similar elements. For instance, in the NLI task, the presence of the word 'not' often serves as a convenient shortcut to identifying contradictions in most training data. However, relying solely on this shortcut without fully understanding the semantic context of the text leads to suboptimal performance when the model encounters a distribution shift in the input data [24], [25]. A focal point of concern has been the need to address lexical bias, denoting the inclination of NLU models to depend on spurious correlations between shortcut words and corresponding labels [26], [27]. It becomes essential to tackle this bias to enhance the robustness and reliability of large language models.

Furthermore, it has been observed that the performance of BERT-like models in the NLI task can be largely explained by relying on fictitious statistical inputs such as the unigrams 'neither', 'did', 'has', and the bigrams 'need not'. These statistical patterns play a significant role in contributing to the models' performance in NLI tasks [4]. Concerning this, the long-tailed phenomenon is suggested as a potential explanation for the fast learning capabilities of NLU models. It involves the use of local mutual information as a measure [28] between feature x and label y resulting in a representation of features that follows a long-tailed distribution in the training set [29]. Equation (1) describes this distribution, where certain words or phrases with high mutual information display strong associations with specific class labels.

$$LMI\,(x, y) = p\,(x, y)\,.\log(\frac{p\,(y\,|\,x)}{p\,(y)}) \tag{1}$$

Here, $(x, y) = \frac{count(x,y)}{|N|}$, $p\,(y\,|\,x) = \frac{count(x,y)}{count(x)}$, $|N|$ is the number of features in the training samples, $count(x, y)$ represents co-occurrence of feature x with label y, $count(x)$ is the total features in the training samples. These NLU models focus mostly on information near the top of the distribution [22], which typically corresponds to non-generalizable shortcut features, by using an interpretation approach to analyze model behavior.

Additionally, during the training process, NLU models frequently detect shortcut features in very early iterations. To discourage the NLU model from producing overconfident predictions for training samples with high shortcut degrees, LTGR (Long-Tailed Distribution Guided Regularizer) is implemented using the knowledge distillation framework. It compels the model to smooth the original probability s_j, as given in (2). The logit value and softmax value of the training sample x_j are derived using the biased teacher model as z_j^T and $\sigma\left(z_j^T\right)$ respectively, where σ is the softmax function.

$$s_j = \frac{\sigma\left(z_j^T\right)_m^{1-b_j}}{\sum_{l=1}^{L}\sigma\left(z_j^T\right)_l^{1-b_j}} \tag{2}$$

Here, L represents the total labels. If $b_j = 0$, then s_j will be as same as $\sigma\left(z_j^T\right)$, means there will be no penalization. If $b_j = 1$, then s_j will have the same value for all L labels. Based on the greater of the shortcut degree measurement of each training sample b_j discouraging the NLU model from making excessively accurate predictions for instances that have a substantial shortcut degree.

Similarly, in the reading comprehension task, models often rely solely on the lexical correlation between the words in the question and the source material, overlooking the design of the reading comprehension task itself [30]. Consequently, the evaluation of LLM-based reading comprehension (RC) models has primarily focused on their ability to handle challenging RC tasks by effectively understanding the semantic relationships between words. However, it has been acknowledged that LLMs often lack sufficient training and knowledge in this regard. To address this issue, artificial adversarial examples are generated to study shortcut learning processes and use adversarial data augmentation to strengthen the models [31]. Mathematically, the model f is evaluated by taking paragraph-question pair (p,q) and outputting the answer $\hat{a}$ to check between the true answer a and the predicted answer $f\,(p,q)$ using the F-1 score s. The *standard accuracy* for a test sample T_{test} is given in (3).

$$Accuracy\,(f) = \frac{1}{T_{test}}\sum\nolimits_{(p,q,a)\in T_{test}} s\,((p, q, a)\,, f) \tag{3}$$

These studies claim to have high success rates for fake adversarial examples, however, it is unclear how well they are still effective for distributions from real-world applications [32]. In the literature, the Semantic Role Labeling combined with Ant Colony Optimization technique is used to generate additional training data for sentiment analysis task [33]. Comparatively, the adversarial evaluation shows that the current models are too resistant to alterations that change semantics. There may be a need for new training model procedures to optimize adversarial evaluation metrics. The adversary function F takes an instance (p, q, a), with model f, and creates a new instance $\left(p', q', a'\right)$. The accuracy of the adversary w.r.t. F is given in (4).

$$Adversary\,(f) = \frac{1}{T_{test}}\sum\nolimits_{(p,q,a)\in T_{test}} s\,(F\,(p, q, a, f)\,, f) \tag{4}$$

The adversarial accuracy estimates the fraction of the time that the model is robustly correct in the face of adversarially chosen changes. The $\left(p', q', a'\right)$ should be close to (p, q, a).

It has been stated that spurious feature-label correlations arise due to biased data distributions in the training data which are frequently exploited as shortcuts by the models [34] and specifically, for any input x and associated label y, a classifier F concentrates to approximate the underlying distribution $p = (y \mid x)$ for every $(x, y) \in S$, where S is the whole input-output space for the task. Data is often taken from a constrained domain, $S_a \in S$. Consider $f_e(.)$ represents a spurious feature extractor. While spurious features have no causal relationship to the label in the general domain S as given in (6) and are therefore useless for prediction there, they can be used efficiently for prediction in the confined domain S_a as in (5).

$$p\left(y \mid f_e\left(x\right)\right) \approx p\left(y \mid x\right), \quad (x, y) \in S_a \tag{5}$$

$$p\left(y \mid f_e\left(x\right)\right) \approx p(y), \quad (x, y) \in S \tag{6}$$

To address data bias in classification problems, it is essential to avoid assigning excessive weight to any specific input feature when constructing a classifier from labeled data. This helps to create a more robust classifier [35]. Regularization is a common technique employed to distribute weights more evenly among features, thereby enhancing robustness. However, since regularization is a general approach, it may not provide tailored resilience specific to the classification problem at hand. To overcome this limitation, researchers have proposed the use of game-theoretic formalization to analyze resilience and prevent the over-weighting of individual features. By adopting this approach, classifiers can be developed that are optimally resilient to feature deletion in a minimax sense [24]. The classifiers created in this work are designed to be resilient to feature deletion in a minimax sense. The goal is to develop classifiers that can maintain their performance even when certain features are removed or altered. The study proposes a method utilizing quadratic programming to construct such classifiers. By optimizing certain objective functions through quadratic programming techniques, the resulting classifiers exhibit robustness against feature deletion, ensuring their durability and effectiveness in various scenarios. The classifiers created by the authors are resistant to feature deletion where a labeled instance $(x_t, y_t)\,(i = 1, .., n)$ along with input feature vectors $x_t \in R^n$ and $y_t \in \{\pm 1\}$. The constant K denotes the maximum number of features that can be deleted for any given point x. When training the classifier, the hinge loss $\sum_t [1 - y_t w.x_t]_+$ is intended to be utilized to measure the objective function to maximize [27]. The classifier may develop resistance to random feature value set to zero, or to test-time feature deletion from the input vectors. Due to the possibility of deleting up to K features from each data vector, a classifier must be developed that minimizes the worst-case hinge loss. Equation 7 provides the worst-case hinge loss for example t in this situation.

$$h^{wc}\left(w, y_t x_t\right) = \max\left[1 - y_t w.\left(x_t^{\circ}\left(1 - \alpha_t\right)\right)\right]_+ \quad s.t.\ \alpha_t \in \{0, 1\}, \sum_j \alpha_{tj} = K \tag{7}$$

The maximization in this case is across all permissible assignments to α_t, where α_{tj} stands for the jth element of α_t that equals 1 if the jth feature of x_t is removed (the symbol $^{\circ}$ dot represents the element-wise multiplication operation).

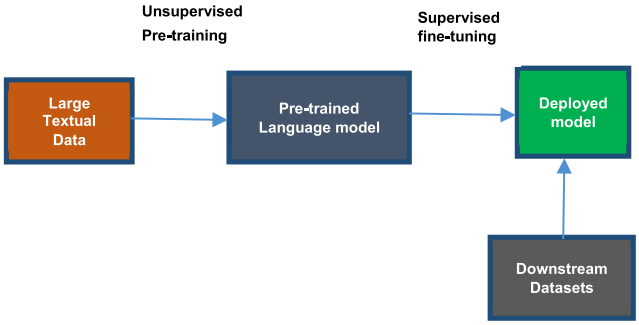

FIGURE 2. Pre-training and fine-tuning training mechanism of large language model.

Shortcut learning in the context of Large Language Models (LLMs) has become a focal point of investigation due to its potential to impact the reliability and generalization capabilities of these models. The problem of learning shortcuts in Large Language models is driven by a variety of factors in the training process as illustrated in Fig. 2. This literature review seeks to dissect and analyze these factors, shedding light on the complexities of shortcut learning in LLMs. Special emphasis is placed on two critical dimensions: dataset biases during training, and the robustness challenges faced by LLMs.

A. DATASET BIASES: TRAINING IMBALANCED DATASETS

The presence of annotation artifacts and collection artifacts in training data can contribute to shortcut learning in large language models. Annotation artifacts refer to biases or errors introduced during the data labeling or annotation process, while col-lection artifacts are biases introduced due to the specific way data is collected or sampled. Large language models have the potential to pick up biases from a variety of textual sources, including biased or unreliable material from the web. This can lead to the presence of collection artifacts in the training data. When multiple crowd workers are used to annotate data, biases, errors, or consistency issues may occur. These issues are known as annotation artifacts.

The studies revealed how LLMs take advantage of annotation artifacts in datasets to achieve high accuracy even when the sentence is not understood [35]. The use of heuristics by annotators to quickly and effectively create hypotheses could be one justification for the emergence and comparative reliability of such dataset artifacts [36]. These artifacts can be unintentional, resulting from varying interpretations, subjective judgments, or inherent biases of the annotators. Annotation artifacts can harm the reliability and accuracy of labeled data used to train large language models. These biases may manifest in several ways, such as the use of biased phrases, the underrepresentation of particular demographics

or viewpoints, or the spread of stereotypes. Language biases need to be quantified and understood since they can serve to support the psychological status of certain demographic groups [37].

Furthermore, it has been shown that supervised linguistic interpretation models depend substantially on dataset artifacts, specifically the inclination of some words to function as prototype hypernyms [38]. The article highlights how modern visual question-and-answer systems take advantage of annotation biases in the dataset [39]. It was revealed that sophisticated models for referencing expression identification operate well even in the absence of text input [40]. These results are consistent with prior studies and strongly imply that supervised models will make use of data shortcuts to manipulate the benchmark if any exist [41].

The authors demonstrate that even when two phrases have highly different meanings, such as ''The famous and snobbish cat isn't nastier than the dog in a white dress and glasses'' and ''The dog in a white dress and glasses isn't nastier than the famous and snobbish cat'', for sentences with significant word overlap, models can still be employed to predict the entailment relation [42] which leads to overlap bias. Paraphrase identification has also had similar failures. A dataset with a substantial lexical overlap of paraphrase and non-paraphrase pairs has been created. Simple BOW models are poor at capturing non-local contextual information, as shown by their inability to learn from such training datasets [43]. In numerous contexts over several decades, gender bias in language has been investigated [17].

A new NLI test set was developed to demonstrate the shortcomings of recent models in conclusions that require lexical and contextual knowledge. The modern NLI systems have a limited capacity for generalization and frequently miss out on simple inferences that require lexical and contextual knowledge [44]. The goal of typical de-biasing techniques is to eliminate biased features from the learned representation. However, biased features in textual data frequently mix superficial cues with useful semantic information, so eliminating them might have a significant negative impact on forecasting accuracy [45]. Nevertheless, we are certain that the bias won't affect prediction, the research demonstrates the inadequacy of current bias reduction techniques [46]. Some methods use ''hard'' examples to account for category shifts given biased features that cannot be accurately predicted using only biased features, as op-posed to de-biasing the data representation [47].

It has long been understood that training dataset biases provide a challenge for machine learning algorithms. Even ''big data'' have a high size and range exhibit biases, therefore many enormous real-life datasets typically provide plenty of shortcut opportunities. Even if eliminating bias is crucial, finding techniques to stop models from exploiting well-known biases may permit us to continue to utilize current datasets and to update the methodologies as the knowledge of the biases is wish to evade grows.

B. ROBUSTNESS ISSUES IN LLMS

Large language models (LLMs) can exhibit vulnerabilities in terms of robustness, including sensitivity to input perturbations and susceptibility to adversarial attacks. LLMs can be highly sensitive to minor changes or perturbations in the input, resulting in significant variations in their outputs. The authors gave a neural machine translation (NMT) example where a single character change in the input caused the model to provide a worse translation [48]. Translation Quality (TQ) is given as a metric for measuring quality, such as BLEU [49]. And robustness is measured with TQ as given in Equation 8. Suppose an NMT model M was to translate an input x to y' and its perturbed version x_δ to y'_δ, the translation quality (TQ) on these datasets would be evaluated in comparison to reference translations y : $TQ\left(y', y\right)$ and $TQ\left(y'_\delta, y\right)$ and robustness is defined as follows in (8).

$$ROBUST\left(M \mid x, y, \delta\right) = \frac{TQ\left(y'_\delta, y\right)}{TQ\left(y', y\right)} \tag{8}$$

Using the dataset (x, y), $ROBUST\left(M \mid x, y, \delta\right)$ comes less than 1 indicates that the model $M's$ translation quality degrades in the presence of perturbation, whereas $ROBUST\left(M \mid x, y, \delta\right) = 1$ shows that the model M is robust to perturbation δ.

The robustness issue in LLMs has been investigated concerning adversarial attacks, where purposely crafted input examples are used to manipulate the model's behavior, resulting in inaccurate or unexpected outputs. The authors specifically focus on evaluating reading comprehension systems, including LLM-based models, using adversarial instances that demonstrate the models' sensitivity to even minor changes in input [31]. They highlight that small perturbations can significantly alter the meaning of a document, and despite attempts to enhance robustness through adversarial training, improvements are not observed. This emphasizes the resilience and diversity of adversarial examples [50]. Additionally, to generate adversarial samples that deceive well-trained sentiment analysis and textual entailment models, a black-box population-based optimization approach is employed, which ensures the creation of semantically and syntactically similar adversarial instances. To quantitatively assess the vulnerability of neural networks to adversarial attacks, the authors introduced a robustness score [51]. They also presented an evaluation framework that encompasses both white-box and black-box attack scenarios, providing a comprehensive assessment of the network's resilience.

The authors provide evidence that by emphasizing term/phrase level matching rather than compositionality learning, NLI models can attain a high level of accuracy. This approach has the potential to alleviate the brittleness commonly observed in LLMs, enhancing their robustness and performance [52]. By changing the MNLI development set, the authors created datasets that demonstrate bias [53]. To ensure the uniformity of the input text distribution, it [54] additionally mines new crowds out of the existing

MNLI dataset in their evaluation in addition to using the datasets from the existing study [53]. Data augmentation and enhancement techniques were used to mitigate bias in the dataset. However, the authors found that despite these efforts, the approach was not sufficient to eliminate model bias. To ensure reliable and consistent performance in real-world circumstances, it is essential to comprehend and mitigate these difficulties. The following section discusses the origin of dataset bias followed by sections on detection methods and mitigation strategies of dataset bias to avoid shortcut learning.

IV. DATASET-BIAS DETECTION FOR LANGUAGE MODELS

In recent years, numerous metrics have been developed to measure bias in language models. There are two types of bias measures, based on the extent that the language model has been modified for downstream tasks: intrinsic and extrinsic measures [55], [56]. The earlier, which relies on the Masked Language Model (MLM), assesses the bias present in pre-trained language models through several methods for estimating the probability of sensitive features. The likelihood of true as well as false positives is then calculated to see how bias spreads in downstream tasks. Next, this possible bias is removed for the downstream task by employing consistent regularisation to differentiate between biased and unbiased sample illustrations or probability distributions [57]. There has been little research into how bias influences the efficiency of pre-trained language models (PLMs), although PLM biases have been removed and decreased in numerous research, a field where bias is pervasive. How the corresponding bias changes as PLMs are trained and modified in the field is unknown. Furthermore, there is still scope for research into how to reduce the bias in PLMs. Some techniques can help detect bias in the dataset for LLMs.

•Bias in NLP systems can be introduced by training data, resources, pre-trained models such as word embeddings, and algorithms, which might result in predictions that are skewed towards one gender or another. The semantic meanings of words are captured by word embeddings, which are numerical representations of words. Any biases in the data can be found by looking at the word embeddings in the training set. For instance, the language model may be biased if some terms are more frequently linked with one demographic group than another. In addition to proposing a method for debiasing word embeddings, the authors suggested the idea of employing word embedding analysis to identify bias in LLMs [17]. A technique was used for quantifying stereotype bias in word embeddings and how it might be applied to reduce bias in LLMs [58].

•Biases in the emotional tone of the training data can be found through sentiment analysis. This entails examining the text data sentiments in the training dataset, such as positive or negative sentiment toward particular demographic groups and can assist in locating any causes of bias. Analyzing demographic disparities entails evaluating how well the language model performs for various demographic groups. This can reveal any performance gaps or a bias in the model's predictions in favor of particular groups [59].

•In counterfactual analysis, biased sources are removed from the training data. This may entail producing artificial data to represent various demographic groups or altering existing data to correct for biases. The model can be retrained using the updated data and then tested to see if bias has been decreased. Indirect bias in LLMs, which can happen when specific racial or ethnic groups are overrepresented in the training data, was addressed in the research [60].

•The fairness of the language model's predictions for various demographic groups can be evaluated using fairness metrics. A fairness indicator, for instance, can quantify the percentage of accurate forecasts for various demographic groups and reveal any bias or discrepancies in the model's predictions. According to the study, a fairness indicator dubbed ''disparate mistreatment'' can be used to assess how fair LLMs are to certain demographic groups [61].

Current research on gender bias in NLP has concentrated on quantifying bias using psychological testing, performance variations between genders for different tasks, and the geometry of vector spaces. The Implicit Association Test (IAT) is a tool used in psychology to assess unconscious gender bias in people. It measures how quickly and accurately people categorize words as belonging to two ideas that are similar or dis-similar [5]. The authors used the Word Embedding Association Test (WEAT) to quantify bias in word embeddings utilizing the IAT's central idea of determining gender bias through differences in the strength of concept association [5].

Several bias metrics have since been created, including the relational inner product association (RIPA) [62], mean average cosine similarity (MAC) [63], relative negative norm distance (RND) [64], relative negative sentiment bias (RNSB) [65], and a kNN-based metric [46]. Recent Language Models have made these metrics ineffective or require modification to work with state-of-art word embeddings. To detect bias in recent context-aware language models, in extrinsic approaches, bias detection algorithms analyze the performance difference for terms related to two separate target groups in downstream tasks like text classification [56]. The study focuses on identifying a bias subspace inside intrinsic bias detection methods, which examines the influence of the context and modifies direct bias to work for ELMo representations of occupation terms [66]. It also monitors a two-dimensional gender subspace, where bias is visualized by projecting ELMo embeddings of occupation terms [67].

The methods for detecting intrinsic bias make use of several word association tests and can be further separated into LM (language models) and WEAT-based (word embedding association test) techniques. The WEAT is a statistical test based on data and results from the IAT. The authors agree with the existence of human biases in GloVe and Word2Vec embeddings that were discovered during IAT tests. Second, WEAT-based tests let us compare bias in language models based purely on embeddings and predictions. Finally, WEAT-based tests, which have received the largest research

contributions, need to be incorporated. It calculates the distances between word representations in sets of target and attribute words. The strategies for identifying and reducing bias in machine learning models, including LLMs, were surveyed in recent research studies [59]. However, it becomes essential to examine several methods to mitigate dataset bias in Language Models (LLMs).

V. STRATEGIES FOR MITIGATING DATASET BIAS IN LANGUAGE MODELS

Especially in applications that have practical implications, such as natural language processing (NLP) for healthcare, education, or finance, mitigating dataset bias is crucial while developing LLMs. It is crucial to properly assess the training data and make sure that it is representative of the target distribution to mitigate the negative impacts of dataset bias on shortcut learning in language models. Different strategies and techniques are used to mitigate the effects of bias in the training data while minimizing dataset bias in LLMs. The objective is to create a more reliable and unbiased LLM that can deliver more accurate and equitable predictions or outputs across various demographic groups.

Recent research has offered a variety of countermeasures to language model shortcut learning caused by dataset bias:

•By performing various transformations or adjustments on the initial dataset, data augmentation includes producing extra training data. Enhancing the diversity of the training data and ensuring that the model is exposed to a larger range of input patterns, can aid in the reduction of bias. The authors propose a method called ''back-translation'' that generates augmented data by translating sentences into a different language and then back to the original language on the IMDb text classification dataset which improves the generalization performance of models and reduces biases [68].

•By using information from a relevant source domain, domain adaptation strategies seek to enhance the effectiveness of a model on a target domain. Enabling the model to generalize to new domains and input patterns, can help to reduce the impacts of bias. The study explores such techniques as domain adversarial training and self-training to enable models to generalize to new domains and reduce the impact of bias [69].

•Regularization approaches can aid in lowering a model's sensitivity to erratic or pointless input elements. Regularisation can assist in preventing the model from overfitting to biased input patterns by penalizing excessively complex or specific patterns in the training data. The study investigates the effects of techniques such as dropout, weight decay, and input noise injection on improving model robustness by penalizing the influence of erratic or irrelevant input elements. It explores the Meta-learning method that learns to weight training samples based on gradient directions [70].

•To make sure that the model generates output that is fair and unbiased concerning delicate traits like race or gender, fairness limits might be included in the training process. The research shows that it is possible to reduce bias while keeping critical contextual information for high-fidelity text generation [71].

•Bias can be reduced by selecting the training data carefully. This can entail selecting samples from a wider range of sources or meticulously selecting the dataset to guarantee that it accurately reflects the target distribution. The authors provide a strategy for gathering training data from a variety of sources, assuring a more diverse representation of insights and lowering the possibility of bias in the final models [72].

Several studies follow the above-mentioned approaches in different applications of natural language processing to avoid dataset biases. The study proposes to use adversarial learning to reduce undesired biases in machine learning models [73] and it explores that a neural network's latent representation can be modified using an adversarial training technique to exclude information about the sensitive attribute. The authors emphasize that biases existing in the training data might be amplified and perpetuated by machine learning models, including LLMs. The debiasing model is trained to identify and eliminate the bias signals that are present in the input data, while the classifier model is trained on the original dataset to carry out the classification task. To reduce the classifier's capacity to predict the biased attribute, the debiasing model is trained as an adversary to the classifier.

The authors demonstrate that the performance of text classification algorithms can be significantly impacted by dataset biases, such as label imbalance or skewed data distributions, the authors point out. Biases in the data can produce biased model predictions and have a detrimental effect on the model's overall efficacy [33]. The authors suggest an approach termed ''instance weighting with majority class correction'' (IW-MCC), which modifies the weight of each training instance based on its class distribution. To balance the influence of majority and minority classes, they specifically provide a strategy for modifying the instance weights in the training data. The following section compares different bias measures for shortcut learning in language models.

VI. COMPARATIVE ANALYSIS OF BIAS MEASURES FOR SHORTCUT LEARNING IN LARGE LANGUAGE MODELS

Machine learning uses data to create models that can evaluate the categories and attributes of new data. However, training data often includes biases in areas that we would prefer not to use for decision-making. Machine learning creates models that are accurate to training data, which might lead to the perpetuation of these unwanted biases. While the ability to generate coherent text is becoming more and more useful, it also encourages models to internalize social biases found in training corpora. Thus, there has been a lot of research interest in examining the social impact and fairness of text produced using language models [74]. A range of cultural connections and unfavorable social biases can be detected by NLP algorithms. Numerous NLP tasks showed comprehensive imbalances, including gender biases in word embedding [58] and [67], biases in sentiment analysis [74], and linguistic generational biases due to demographics [65].

TABLE 1. NLP subjects and bias measures.

Reference	Subject	Criteria or Measure	Limitations
[76]	Co-reference resolution	De-biasing Techniques for Gender Bias through covering a wide range of names, pronouns, and contexts	Generalization-Challenges, Contextual Complexity
[17]	Gender Bias	Generating counterfactual examples by altering gender-associated words to mitigate bias in the embeddings	Language-Specific: May not be applied to other than the English language
[33]	Stop-word Biases	Spurious features through spurious correlations, regularization techniques that discourage over-reliance on specific features	Limited to text classification tasks
[22]	Lexical bias	Robust Training Data, Task-Specific Training	Partial Elimination, Generalization Challenges
[28]	De-biasing fact verification	Fact-checking classification through claim-evidence reasoning, Adversarial Training	Dataset Dependency, Generalization Challenges
[14], [77]	Racial Bias	In diverse and Balanced Datasets, annotators avoid biases and consider sentiment independently of gender or race	Annotation Subjectivity
[13]	Sentiment Bias	Data augmentation, fine-tuning with a larger corpus, Adversarial Training	Limited to gender terms not to racial or sentiment analysis
[78], [79], [80]	Socioeconomic Bias	Data Augmentation with a Range of Socioeconomic Contexts	Challenges in capturing the nuanced biases related to socioeconomic factors
[17], [81]	Political Bias	Adversarial training to reduce the model's reliance on politically biased features	Difficulties in defining and capturing the complex nature of political biases
[73]	Cultural Bias	Ensuring diverse cultural perspectives in the training data	Capturing all cultural biases or avoiding overgeneralizations might be challenging.
[5]	Ideological Bias	Including diverse ideological perspectives in the training data	Difficulties in encompassing the full spectrum of ideological biases

In the ideal scenario, we would be able to create a model that accurately captures the generalizations from the data that are required for completing a task but does not discriminate in a manner that the models deem unjust. Several useful measurements for fairness have been defined in the work of training machine learning systems to output fair decisions: Conditional Demographic Parity, Counterfactual Fairness, Equalized Odds, and more. The context of adversarial debiasing is used to investigate these fairness measures [73]. In another study, counterfactual data augmentation was suggested as a solution to occupation-specific gender biases, and it was discovered that it can significantly better maintain model performance than debiasing word embeddings, notably in language modeling [75]. There are some points for comparative analysis of mitigating strategies to deal with shortcut learning in large language models by different authors as shown in Table 1.

VII. CONCLUSION AND FUTURE DIRECTION

It is concluded that dataset bias can have a considerable impact on the performance of language models, especially when it comes to shortcut learning. It is mentioned that when a language model learns to rely on overrepresented input patterns or features in the training data, it may lead to poor generalization to new inputs, particularly those that do not correspond to the biased patterns. There are several pieces of literature discussed to reduce the effects of dataset bias on shortcut learning, it was highlighted that the training data has to be checked carefully and appropriate mitigation approaches have to be used such as data augmentation, domain adaption, model regularization, data selection, and fairness constraints. In situations where shortcut learning may be particularly harmful, such as in critical applications like healthcare, finance, and law enforcement, it is possible to mitigate the effects of dataset bias and increase the accuracy and fairness of language models by doing this. Due to dataset bias in LLMs, the review on shortcut learning was necessary to focus on existing research that is generating more effective mitigation strategies that can handle the problems with bias and shortcut learning in a range of contexts.

Some possible research directions could deal with the research questions stated in Section I, reduce the effect of dataset bias on shortcut learning in LLMs, and enhance the accuracy and fairness of language models in several applications through the following strategies.

A. DATA AUGMENTATION METHODS

It is possible to reduce the impact of dataset biases on shortcut learning in LLMs by developing more advanced data augmentation techniques that produce synthetic training data that simulate the distribution of real-world inputs. The models can be exposed to a wider range of situations by supplementing the training data with synthetic instances that capture the diversity and complexity of real-world inputs, which reduces

their reliance on shortcuts generated from biased or limited training data. The models can generalize more effectively and counteract the negative effects of dataset biases when using these enhanced data, which can give a more accurate representation of the underlying input distribution. To enable the models to learn robust representations and prevent shortcut learning based just on biases or surface-level patterns, it is necessary to produce augmented data that includes the variations, nuances, and unusual situations present in the real-world inputs.

In this study, EDA (Easy Data Augmentation) [82], a straightforward yet efficient data augmentation technique for text classification tasks, is introduced. To create augmented data, EDA uses four different text transformations: synonym replacement, random insertion, random swap, and random deletion. This successfully expands the size and diversity of the training set. Data augmentation for low-resource neural machine translation is the primary focus of other research [83]. It investigates several augmentation methods, including back-translation, which creates fictitious source-target sentence pairs by applying a reverse translation model, thereby enhancing the training data and better representing the target distribution. Additionally, the authors offer a data augmentation technique created especially for BERT-based text classification models [84]. It suggests several augmentation methods, including synonym substitution, random insertion, random swapping, and random deletion, to produce varied training examples that enhance model performance and reduce shortcut learning.

B. REGULARIZATION METHODS

It is possible to reduce the impact of dataset biases on shortcut learning in LLMs by improving model regularization techniques to prevent biased patterns from being overfitted without sacrificing precision. Researchers started hoping to encourage fair and robust learning while keeping high precision by incorporating regularization approaches that specifically discourage the models from depending on false correlation and biased patterns. To reduce the effects of dataset biases, the authors suggest a fairness regularization strategy that penalizes models for producing predictions based on protected and biased features [85]. Furthermore, a framework for classification that is fairness-aware and uses regularization methods to enforce fairness requirements is proposed. The strategy tries to reduce biases and avoid overfitting biased patterns by introducing fairness restrictions during model training [70].

C. 7.3. EXPLAINABLE AI TOOLS

An important research direction to reduce dataset bias and accelerate learning in LLMs is the creation of explainable AI tools to aid in identifying and diagnosing bias in language models and devising efficient mitigation techniques. These techniques may help academics and practitioners better understand and address bias-related problems by supplying insights into the decision-making process of the model and locating potential sources of bias. To identify and diagnose shortcut learning in transformer-based language models, an explainable AI tool called ''Probing Learning Explainers'' (PROLEX) is designed [86]. Researchers can find and examine the model's shortcuts through PROLEX's fine-grained explanations of model predictions. Based on the knowledge gathered from the explainable tool, it also suggests methods for reducing shortcut learning. The ''DIAG'' diagnostic framework is also introduced, which makes it easier to recognize and measure shortcut learning. It offers insight into creating stronger regularization approaches to mitigate the shortcomings and difficulties related to shortcut learning [87].

These future directions highlight the significance of different approaches in uncovering and diagnosing bias in language models and designing effective mitigation mechanisms. By using these techniques, researchers can gain a better understanding of shortcut learning, identify biases, and develop strategies to reduce dataset bias and improve the overall fairness and accuracy of LLMs.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.
[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ''RoBERTa: A robustly optimized BERT pretraining approach,'' 2019, *arXiv:1907.11692*.
[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, ''Language models are few-shot learners,'' in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.
[4] T. Niven and H.-Y. Kao, ''Probing neural network comprehension of natural language arguments,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4658–4664, doi: 10.18653/v1/p19-1459.
[5] A. Caliskan, J. J. Bryson, and A. Narayanan, ''Semantics derived automatically from language corpora contain human-like biases,'' *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 2017.
[6] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, ''The measure and mismeasure of fairness,'' 2018, *arXiv:1808.00023*.
[7] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, ''Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping,'' 2019, *arXiv:2002.06305*.
[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, and P. Schuh, ''PaLM: Scaling language modeling with pathways,'' *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.
[9] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu, ''Rethinking softmax cross-entropy loss for adversarial robustness,'' 2019, *arXiv:1905.10626*.
[10] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou, and H. W. Hon, ''UniLMv2: Pseudo-masked language models for unified language model pre-training,'' in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 642–652. [Online]. Available: https://proceedings.mlr.press/v119/bao20a.html
[11] B. Giovanola and S. Tiribelli, ''Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms,'' *AI Soc.*, vol. 38, no. 2, pp. 549–563, Apr. 2023, doi: 10.1007/s00146-022-01455-6.
[12] S. L. Blodgett, H. D. Iii, and H. Wallach, ''Language (technology) is power: A critical survey of 'bias' in NLP,'' 2020, *arXiv:2005.14050*.
[13] J. Ho Park, J. Shin, and P. Fung, ''Reducing gender bias in abusive language detection,'' 2018, *arXiv:1808.07231*.
[14] S. Kiritchenko and S. M. Mohammad, ''Examining gender and race bias in two hundred sentiment analysis systems,'' 2018, *arXiv:1805.04508*.
[15] M. Du, F. He, N. Zou, D. Tao, and X. Hu, ''Shortcut learning of large language models in natural language understanding,'' *Commun. ACM*, vol. 67, no. 1, pp. 110–120, 2023.

[16] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. Int. Conf. Softw. Eng.*, 2018, pp. 1–7, doi: 10.1145/3194770.3194776.
[17] A. T. K. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
[18] K. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19000–19015.
[19] L. J. Cohen, "On the psychology of prediction: Whose is the fallacy?" *Cognition*, vol. 7, no. 4, pp. 385–407, Jan. 1979, doi: 10.1016/0010-0277(79)90023-4.
[20] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," 2017, *arXiv:1710.03184*.
[21] T. Niven and H.-Y. Kao, "NLITrans at SemEval-2018 task 12: Transfer of semantic knowledge for argument comprehension," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1099–1103, doi: 10.18653/v1/s18-1185.
[22] M. Du, V. Manjunatha, R. Jain, R. Deshpande, F. Dernoncourt, J. Gu, T. Sun, and X. Hu, "Towards interpreting and mitigating shortcut learning behavior of NLU models," in *Proc. NAACL*, 2021, pp. 915–929, doi: 10.18653/v1/2021.naacl-main.71.
[23] R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: 10.1038/s42256-020-00257-z.
[24] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proc. ACM Int. Conf.*, vol. 148, 2006, pp. 353–360, doi: 10.1145/1143844.1143889.
[25] V. Dogra, A. Singh, S. Verma, A. Alharbi, and W. Alosaimi, "Event study: Advanced machine learning and statistical technique for analyzing sustainability in banking stocks," *Mathematics*, vol. 9, no. 24, p. 3319, 2021.
[26] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" 2019, *arXiv:1909.01066*.
[27] C. Gentile and M. K. Warmuth, "Linear Hinge loss and average margin," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 11, 1998, pp. 1–7.
[28] T. Schuster, D. J. Shah, Y. J. S. Yeo, D. Filizzola, E. Santus, and R. Barzilay, "Towards debiasing fact verification models," in *Proc. Conf. Empir. Methods Nat. Lang. Process., 9th Int. Jt. Conf. Nat. Lang. Process.*, 2019, pp. 3419–3425, doi: 10.18653/v1/d19-1341.
[29] S. Evert, "The statistics of word cooccurrences word pairs and collocations," Ph.D. dissertation, Stuttgart Univ., Germany, 2005.
[30] S. Kwon, C. Kang, J. Han, and J. Choi, "Why do masked neural language models still need common sense knowledge?" 2019, *arXiv:1911.03024*.
[31] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2021–2031, doi: 10.18653/v1/d17-1215.
[32] J. X. Morris, E. Lifland, J. Lanchantin, Y. Ji, and Y. Qi, "Reevaluating adversarial examples in natural language," in *Proc. Find. Assoc. Comput. Linguist. Find. ACL EMNLP*, 2020, pp. 3829–3839, doi: 10.18653/v1/2020.findings-emnlp.341.
[33] A. Liusie, V. Raina, V. Raina, and M. Gales, "Analyzing biases to spurious correlations in text classification tasks," in *Proc. 2nd Conf. Asia–Pacific Chapter Association Comput. Linguistics, 12th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2022, pp. 78–84.
[34] V. Dogra, F. S. Alharithi, R. M. Álvarez, A. Singh, and A. M. Qahtani, "NLP-based application for analyzing private and public banks stocks reaction to news events in the Indian stock exchange," *Systems*, vol. 10, no. 6, p. 233, Nov. 2022, doi: 10.3390/systems10060233.
[35] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 107–112, doi: 10.18653/v1/n18-2017.
[36] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," 2016, *arXiv:1606.01933*.
[37] E. Sapir, *Selected Writings of Edward Sapir in Language, Culture and Personality*. Berkeley, CA, USA: Univ. California Press, 2021.
[38] O. Levy, S. Remus, C. Biemann, and I. Dagan, "Do supervised distributional methods really learn lexical inference relations?" in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Conf.*, 2015, pp. 970–976, doi: 10.3115/v1/n15-1098.
[39] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.
[40] V. Cirik, L.-P. Morency, and T. Berg-Kirkpatrick, "Visual referring expression recognition: What do systems actually learn?" 2018, *arXiv:1805.11818*.
[41] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," 2017, *arXiv:1705.02364*.
[42] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3428–3448, doi: 10.18653/v1/p19-1334.
[43] Y. Zhang, J. Baldridge, and L. He, "PAWS: Paraphrase adversaries from word scrambling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Human Lang. Technol.*, 2019, vol. 1, no. 2, pp. 1298–1308.
[44] M. Glockner, V. Shwartz, and Y. Goldberg, "Breaking NLI systems with sentences that require simple lexical inferences," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 650–655, doi: 10.18653/v1/p18-2103.
[45] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 12907–12929.
[46] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," in *Proc. Conf. Technol., Amer. Chapter Assoc. Comput. Linguistics, Human Lang.*, 2019, pp. 609–614.
[47] H. He, S. Zha, and H. Wang, "Unlearn dataset bias in natural language inference by fitting the residual," in *Proc. 2nd Workshop Deep Learn. Approaches Low-Resource NLP*, 2019, pp. 132–142, doi: 10.18653/v1/d19-6115.
[48] X. Niu, P. Mathur, G. Dinu, and Y. Al-onaizan, "Evaluating robustness to input perturbations for neural machine translation," 2020, *arXiv:2005.00580*.
[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.
[50] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," in *Proc. EMNLP*, 2018, pp. 2890–2896.
[51] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[52] Y. Nie, Y. Wang, and M. Bansal, "Analyzing compositionality-sensitivity of NLI models," in *Proc. 33rd Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, 2019, pp. 6867–6874, doi: 10.1609/aaai.v33i01.33016867.
[53] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig, "Stress test evaluation for natural language inference," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2340–2353.
[54] X. Zhou and M. Bansal, "Towards robustifying NLI models against lexical dataset biases," 2020, *arXiv:2005.04732*.
[55] P. Delobelle, E. Tokpo, T. Calders, and B. Berendt, "Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2022, pp. 1693–1706.
[56] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams, "Multidimensional gender bias classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 314–331.
[57] K. Anoop, M. P. Gangan, P. Deepak, and V. L. Lajish, "Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias," in *Responsible Data Science*. Singapore: Springer, 2022, pp. 13–45.
[58] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Quantifying and reducing stereotypes in word embeddings," 2016, *arXiv:1606.06121*.
[59] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. IEEE Workshop Adv. Robot. Social Impacts (ARSO)*, Mar. 2017, pp. 1–7.
[60] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowl. Inf. Syst.*, vol. 54, no. 1, pp. 95–122, Jan. 2018.
[61] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1171–1180.
[62] K. Ethayarajh, D. Duvenaud, and G. Hirst, "Understanding undesirable word embedding associations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 1696–1705.

[63] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, "Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Lingistics*, 2019, pp. 615–621.
[64] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 16, pp. 3635–3644, Apr. 2018.
[65] C. Sweeney and M. Najafian, "A transparent framework for evaluating unintended demographic bias in word embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1662–1667.
[66] W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2021, pp. 122–133.
[67] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating gender bias in natural language processing: Literature review," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1630–1640.
[68] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6256–6268.
[69] D. Saunders, "Domain adaptation for neural machine translation," Doctoral dissertation, Univ. Cambridge, Cambridge, U.K., 2021.
[70] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, 2017, pp. 962–970. [Online]. Available: https://proceedings.mlr.press/v80/ren18a.html
[71] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 6565–6576.
[72] D. Hovy and S. Prabhumoye, "Five sources of bias in natural language processing," *Lang. Linguistics Compass*, vol. 15, no. 8, pp. 1–19, Aug. 2021, doi: 10.1111/lnc3.12432.
[73] B. Hu Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," 2018, *arXiv:1801.07593*.
[74] P. S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli, "Reducing sentiment bias in language models via counterfactual evaluation," in *Proc. Find. Assoc. Comput. Linguist. Find. ACL EMNLP*, 2020, pp. 65–83, doi: 10.18653/v1/2020.findings-emnlp.7.
[75] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," in *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*. Switzerland: Springer, 2020, pp. 189–202.
[76] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the GAP: A balanced corpus of gendered ambiguous pronouns," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 605–617, Dec. 2018, doi: 10.1162/tacl_a_00240.
[77] D. Card, S. Gabriel, N. A. Smith, P. G. Allen, and C. Science, "The risk of racial bias in hate speech detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1668–1678.
[78] K. Inui, J. Jiang, V. Ng, and X. Wan, "Measure country-level socio-economic indicators with streaming news: An empirical study," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1249–1254.
[79] H. Daumé III, H. Wallach, S. L. Blodgett, and S. Barocas, "Mitigating socioeconomic bias in language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020.
[80] H. Viswanath and T. Zhang, "FairPy: A toolkit for evaluation of social biases and their mitigation in large language models," 2023, *arXiv:2302.05508*.
[81] A. M. Davani, M. Atari, B. Kennedy, and M. Dehghani, "Hate speech classifiers learn normative social stereotypes," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 300–319, Mar. 2023, doi: 10.1162/tacl_a_00550.
[82] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*.
[83] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," 2017, *arXiv:1705.00440*.
[84] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional BERT contextual augmentation," in *Proc. 19th Int. Conf. (ICCS)*, 2019, pp. 84–95.
[85] L. Dixon, L. Li, J. Sorensen, J. Thain, and N. Vasserman, "Mitigating bias in natural language processing models: Interventions and implications," in *Proc. Conf. Fairness, Accountability, Transparency (FAccT)*, Barcelona Spain, 2020.
[86] M. Jain, S. Wallace, E. Singh, and S. Gardner, "Explaining and mitigating shortcut learning in transformer language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2021.
[87] O. Zhang, C. Bengio, S. Hardt, M. Recht, and B. Vinyals, "Towards a comprehensive understanding of shortcut learning in deep neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2020.

VARUN DOGRA received the Ph.D. degree computer applications from the School of Computer Science and Engineering, Lovely Professional University, Punjab, India, in 2022.

He is currently an Associate Professor with the School of Computer Science and Engineering, Lovely Professional University. He has published articles in reputed journals indexed in WoS and Scopus. He has more than 15 years of experience in teaching/industry. His research interests include data science, machine learning, natural language processing, computer vision, and FinTech. He is a reviewer of top-cited journals, such as IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING and *Neural Computing and Applications*.

SAHIL VERMA (Senior Member, IEEE) received the B.Tech., M.Tech., and Ph.D. degrees in computer science engineering from Jaipur National University (JNU), Jaipur, India.

He is currently affiliated with Chitkara University, Punjab, India. He has published papers in reputed top-cited journals. He has chaired many sessions at international conferences. His research interests include the IoT, mobile computing, AI, NLP, and blockchain. He is a Reviewer of top-cited journals, such as IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE ACCESS, *Neural Computing and Applications* (Springer), *Human-centric Computing and Information Sciences* (Springer), *Mobile Networks and Applications* (Springer), *Journal of Information Security and Applications* (Elsevier), *Mobile Information Systems* (Hindawi), *International Journal of Communication Systems* (Wiley), and *Security and Communication Networks* (Hindawi).

KAVITA (Senior Member, IEEE) received the B.Tech., M.Tech., and Ph.D. degrees in computer science engineering from Jaipur National University (JNU), Jaipur, India. She is currently affiliated with the Chitkara University, Punjab, India. She has published articles in reputed top-cited journals. She has chaired many sessions at international conferences. Her research interests include cloud computing, mobile ad-hoc networks, the IoT, wireless sensor networks, AI, and NLP. She is a member of ACM and IAENG. She is also an editorial board member of many international journals. She has chaired many sessions at reputed international conferences in abroad and India. She is a Reviewer of many journals, including the *Multimedia Tools and Applications* and *Computers and Electrical Engineering* journal.

MARCIN WOŹNIAK received the M.Sc. degree in applied mathematics from the Silesian University of Technology, Gliwice, Poland, in 2007, and the Ph.D. degree in computational intelligence and the D.Sc. degree in computational intelligence from the Częstochowa University of Technology, Częstochowa, Poland, in 2012 and 2019, respectively.

In 2022, he received the title of Full Professor in industrial informatics and telecommunication. He is currently a Full Professor with the Faculty of Applied Mathematics, Silesian University of Technology. He is a Scientific Supervisor in editions of "The Diamond Grant" and "The Best of the Best" programs for highly talented students from the Polish Ministry of Science and Higher Education. He participated in various scientific projects (as the Lead Investigator, a Scientific Investigator, the Manager, a Participant, and an Advisor) at Polish, Italian, and Lithuanian universities and projects with applied results in the IT industry both funded by the National Centre for Research and Development and abroad. He was a Visiting Researcher with universities in Italy, Sweden, and Germany. He has authored/coauthored more than 200 research papers in international conferences and journals. His current research interests include neural networks and fuzzy logic control systems with their applications together with various aspects of applied computational intelligence accelerated by evolutionary computation and federated learning models. In 2017, he was awarded by the Polish Ministry of Science and Higher Education with a scholarship for an Outstanding Young Scientist. In 2021, he received an award from the Polish Ministry of Science and Higher Education for research achievements. From 2020 to 2023, he was presented among "Top 2% Scientists in the World" by Stanford University for his career achievements. He was an Editorial Board Member or an Editor of *Machine Learning with Applications*, *Sensors*, *Pattern Analysis and Applications*, IEEE ACCESS, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, *Measurement*, *Sustainable Energy Technologies and Assessments*, *Frontiers in Human Neuroscience*, *PeerJ Computer Science*, *International Journal of Distributed Sensor Networks*, *Computational Intelligence and Neuroscience*, and *Journal of Universal Computer Science*, and the Session Chair at various international conferences and symposiums, including IEEE Symposium Series on Computational Intelligence, International Joint Conference on Neural Networks, and IEEE Congress on Evolutionary Computation.

JANA SHAFI is currently with the Department of Computer Science, Prince Sattam bin Abdulaziz University, Saudi Arabia. She has more than eight years of teaching and research experience. Her research interests include online social networks, wearable technology, artificial intelligence, machine learning, deep learning, smart health, and the IoMT.

MUHAMMAD FAZAL IJAZ was a Visiting Guest Professor and an Assistant Professor with tertiary institutes, including Dongguk University; Tecnológico de Monterrey, Campus Mexico City and Guadalajara, Mexico; Sejong University, Seoul, and The University of Melbourne, Australia. He has published numerous research articles in several international peer-reviewed journals, including IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE INTERNET OF THINGS JOURNAL, *Scientific Reports*, *Cancers*, *Human-centric Computing and Information Science*, *Biomedical Signal Processing and Control*, and *Computational Intelligence*. His research interests include human-centered AI, medical image analysis, medical artificial intelligence, the Internet of Things, and data mining.