IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 March 2023, accepted 1 May 2023, date of publication 16 May 2023, date of current version 6 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3276757

SURVEY

Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools

ANDY DONALD[1], APOSTOLOS GALANOPOULOS[2], EDWARD CURRY[1], EMIR MUÑOZ[2], IHSAN ULLAH[1], M. A. WASKOW[1], MACIEJ DABROWSKI[2], AND MANAN KALRA[2]

[1]Insight SFI Research Center for Data Analytics, University of Galway, Lower Dangan, Galway, H91 AEX4, Ireland
[2]Genesys Cloud Services Inc., Galway, H91 AX8R Ireland

Corresponding author: Edward Curry (edward.curry@universityofgalway.ie)

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement Nos SFI/12/RC/2289_P2 and 20/SP/8955 at the Insight SFI Research Centre at the University of Galway. Insight, the SFI Research Centre for Data Analytics is funded by Science Foundation Ireland through the SFI Research Centres Programme.

ABSTRACT With the increase in usage of machine learning models within many different aspects of customer interactions, it has become very clear that bias detection within associated customer interaction datasets has led to a critical focus on issues such as the identification of bias prior to model building, lack of understanding and transparency within models, and ultimately the prevention of biased predictions or classifications. This has never been more important since the introduction of the EU General Data Protection Regulation (GDPR) and the associated rule of ''right of explanation''. In this paper, we survey the state of the art for bias detection, avoidance and mitigation within datasets, and the associated methods and tools available. Our purpose is to establish an understanding of how established customer interaction-based use cases can utilise these techniques. The focus is primarily on tackling the bias in unstructured text data as a pre-process prior to the machine learning model training phase. We hope that this research encourages the further establishment of responsible usage of customer interaction datasets to allow the prevention of bias being introduced into machine learning pipelines and to also allow greater awareness of the potential for further research in this area.

INDEX TERMS Bias detection, machine learning, bias evaluation, explainable AI.

I. INTRODUCTION

After more than ten years of continuous improvement in AI-based deep learning models, researchers have now started investigating the issues arising from AI-based systems [1]. Besides ethical concerns, other issues at the heart of machine learning are critical to handle. Examples are, biased decisions due to bias in the training data and the model; a system wrongly predicting/classifying an object with high confidence; a lack of understanding of how a decision is taken, or what input features were important in this decision [2]. This can lead to legal complications, such as the lack of adherence to the EU General Data Protection Regulation (GDPR)[1] rule of 'right to explanation,' e.g., a bank client whose loan application has been rejected has the right to know why their application was rejected or to understand why a bot/chatbot responded with a specific answer/information to a query.

According to the Cambridge dictionary [3], bias is *the act of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence one's judgment.* In machine learning (ML), bias can be detected in various places of an ML pipeline such as data collection & planning stage, storage, pre-processing, training, and decision-making (metrics) stage. Levitin [4] highlights that as data is collected by humans, *they* decide what to collect and what not. The objective for which the data is collected and its respective planning leads to wrong analysis/conclusions, e.g., which population/features to select and what to label, also called lexical bias [5]. At the learning stage, it is the bias that exists due to the transfer of bias in the model and how much it affects certain groups while proposing a generalised model that will work for all groups in the data [6]. The trained model is evaluated by its performance on the test set using various metrics, where each performance is dependent on

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

[1]https://www.consilium.europa.eu/en/policies/data-protection/#gdpr

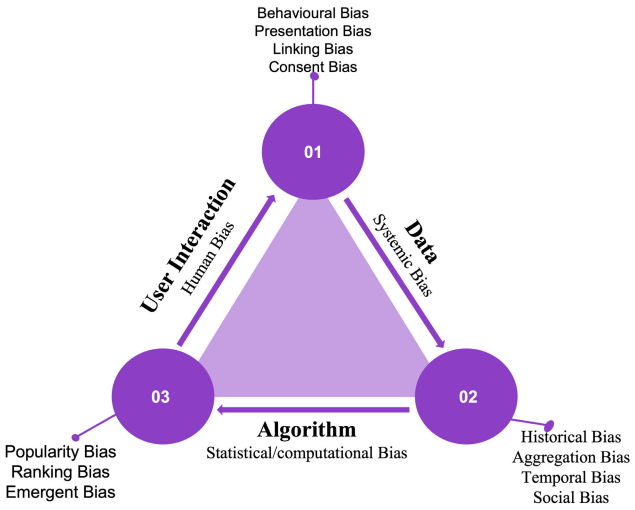


FIGURE 1. An overview of Bias in the ML process.

the application area. For example, customers might be more concerned about false negatives (e.g., being denied a loan when they actually are deserving), and companies care about false positives (e.g., recommending loans to people who dońt pay them back). Other evaluations include — if the model is outputting probabilities — what threshold should be present in order to bolster the decision, and, whether it should be binary or multi-class (e.g., bias, not bias, bias with doubt) to avoid uncertainty in the decisions of these algorithms.

Similarly, fairness in decision-making is another requirement for ML models, which can be achieved by measuring appropriate metrics. For example, Kim et al. [7] enhanced the work of Dwork et al. [8] using computationally bounded awareness, i.e., if the user has one of the metric information, the model can be tested a number of times and analysed for its fairness.

Explainable artificial intelligence (XAI) is a research area that is aiming at the development of machine learning techniques that will enable us to understand, create trust, and give us the ability to manage emerging systems that use AI [2]. The users of this capability to ''explain'' are the data scientists, end-users, company personnel, regulatory authorities or indeed any stakeholder who has a valid remit to ask questions about the decision-making of such systems. Various forms of explainability techniques are used for the detection of bias, e.g., explanation with statistical analysis, and surrogate models.

A detection/debiasing step can be placed in the data processing pipeline that can result in fair and unbiased predictions by trained (fitted) ML models. Under this context, bias in a trained ML model is defined as the existence of any prejudice/favouritism toward an individual or group, based on their inherent or acquired characteristics [9] that came in the data. In contrast, fairness is defined as the absence of any sort of bias.

It is worth highlighting that more often than not, bias, fairness, and explainability overlap; however, they are three distinct fields in machine learning. XAI helps in detecting bias by answering how or why one decision is made, whereas, fairness is a subjective use of ML without favoured or unfair decision-making considering wider aspects of human life [2].

Scope of this paper: Our focus in this paper is mainly on reviewing the tools, techniques, approaches, and methodologies that are used for bias detection, avoidance, and mitigation. More specifically, we focus on pre-processing techniques to tackle bias in unstructured textual data and its usage for ML and/or ML models trained on textual data. This will help in detecting and removing bias before applying any ML algorithm. In order to make our work applicable to real-world scenarios, we will focus on specific use cases for addressing bias in contact-centre environments. The goal is to have specifications for documentation, which can be made available internally and externally describing known biases in each of the solutions.

Next, we present a taxonomy of the different types of bias, based on the recent literature, we describe our use case and present a state-of-the-art review of the existing solutions we can leverage. We will also highlight any XAI techniques that have been used or can be used for highlighting, identifying, and detecting bias in datasets and machine learning models.

The paper is organised as follows: Section II will discuss a few use cases. Section III will discuss some types of biases. It will be followed by Section IV, which is a detailed review of tools and libraries that can be used for bias detection. A comparison of different approaches for the use cases will be presented in Section V. Section VI will highlight the current challenges and some opportunities for research in this domain, and we will conclude the paper in Section VII with final remarks.

II. USE CASES

There are a number of use cases that this paper intends to focus on. These use cases cover scenarios where bias is to be detected as a pre-process in order to define whether or not the dataset analysed is suitable for usage in training an ML model. The ultimate intention here is to present, using bias detection methods in combination with XAI techniques, a clear picture of the biases within the dataset to allow a human decision and possible intervention prior to the ML model-building phase. We determined the following scenarios where bias detection solutions can be put to use.

A. CUSTOMER DATA UPLOADED BY ORGANISATIONS

In many cases, organisations may use external data sources or 3rd party tools for carrying out their business, such as customer data from a CRM (customer relationship management) or ERP (enterprise resource planning) made available to them.

Examples of such data are survey answers that indirectly measure customer loyalty towards a service, and customer profiles from CRMs that are provided ''as-is''. These may

contain hidden biases that organisations may not even be aware of. For example, including survey responses for questions that customers are less likely to answer, sampling them based on customer profiling or only from a particular demography, etc. can lead to inaccurate analyses and false interpretations. If such data is fed directly into the ML pipelines of products/services, it subsequently propagates the biases into the models and later into the predictions [10].

B. AGENTS' BIAS ASSESSMENT

In a contact centre setting, an agent can interact with customers using different media types (e.g., voice, chats, emails). Interactions may contain a transcription of what was said by agents and customers. For example, the transcript can be classified in real time to give feedback to the agent, e.g., on the emotion of the customer [11]. Similarly, the same can be done on understanding the speech/voice of the customer or agent. Since an organisation aims to provide the best customer experience, it would be useful to identify agents (and customers) who present a given bias against the other side. For example, agents that discriminate elderly or behave in a different way with female customers. Clearly, this is unprofessional behaviour that organisations would like to detect and help agents by providing relevant training, for example.

C. LANGUAGE MODELS AND CHATBOTS

Bots, automated AI-based programs designed to mimic human text interaction, are also part of the solution offered by contact-centre solutions [12]. Those are software entities that are usually the first point of contact for a customer, and are charged with collecting basic information by interacting with the client, before potentially involving a human to assist the customer. Tracking and visualising bias in the pipelines for creating or tuning in-house chatbots is a relevant use case. Bias identification at every stage of the creation/tuning process would help organisations make the best-informed decision that benefits their business when wanting to implement chatbots. These bots are also built on top and/or use existing pre-trained models that may also contain bias of some type. Bias discovery and documentation should be considered as well during the documentation of ML pipelines.

III. TYPES OF BIAS

This paper aims to primarily address biases in textual data, that occur during the data collection and dataset construction stages. For instance, racial bias may need to be addressed in some of the review data which consists of inputted text, but it may also need to be addressed in the dataset as a whole if the race categories do not have appropriate representation in the data. Section III-A highlights generic types of biases that might exist in some datasets. Whereas, section III-B focuses on textual datasets specifically, and the types of bias present therein.

A. BIAS IN DATASETS

There are several taxonomies of the types of bias found in datasets, depending on the features of the dataset. The survey in Mehrabi et al. [9] classifies the different types of bias depending on the ML stage at which they appear, e.g., dataset, algorithm, or users, as found in Figure 1. For our purposes, we focus on biases within the dataset as a whole, including:

1) *Measurement bias* arises when data collectors (e.g. researchers, clinicians, participants) use inaccurate methods to measure variables represented in the dataset. This bias is not limited to measurement however, as it also can arise from the way certain features are chosen, utilised, or classified. If customer inquiry trends were being recorded and classified by topic, measurement bias could occur if the classifier was assigning incorrect labels to the topics.
2) *Omitted variable bias* occurs when one or more variables, specifically those important for understanding or accounting for what is represented in the dataset, are left out. When datasets with omitted variables are used to train models, this sort of bias impacts regression models the most [13]. If information about customer use of a chatbot was recorded but the region or time zone information of the customer was not included, this omitted variable bias could lead to inaccurate generalizations of peak customer engagement times in different areas.
3) *Representation bias* refers to skews in the presence or ratio of different demographics in a population, such as a dataset that overrepresents customers within a certain range age, gender, race, or the like. This would lead to issues when a model trained on the dataset is applied to group or individuals who do not fall into that demographic. Representation bias might also arise in the form of semantic representation bias, which can occur in online biographies or while using word embeddings e.g. [14].
4) *Aggregation bias* occurs when false conclusions are drawn about individuals or subgroups based on generalizations of the entire population. It could manifest as modifying a dataset to aggregate the customers by gender and failing to account for subgroup differences such as age or region. This form of bias assumes that mapping from the independent variables to dependent variables is common across various classes/categories of the data [15].
5) *Sampling bias* refers to the non-random sampling of subgroups, and could occur if a dataset meant to be generally representative of the customer population were only sampled from a certain region. It results in lack of generalisation for the trained models since they have not experienced samples from a representative portion of the population [9], [16].
6) *Longitudinal data fallacy* arises when diverse cohorts of temporal data are analysed at a single time point,

as that sort of generalization loses a lot of information that can come overtime [17]. As an example, [18] and [9] highlighted changes in Reddit data where comment lengths were considered to be decreasing in general over time. However, when bulk data was considered, it showed the opposite trend i.e. comment length increased over time when data from cross-sectional snapshot of the population from different years were analysed.

7) *Linking bias* arises when network attributes misrepresent the true behaviour of users. According to [19], the features such as connections, interactions, or activity obtained to form a network might result in variance in attributes. This difference or variance might be considered as behavioral biases.
8) *Unintended bias* is a broad type as it can be in various forms that can affect ML models and their results. For example, Nozza et al. [20] worked on unintended bias detection in Misogyny Detection models. The bias in these models is in the form of detecting and scoring high specific words e.g. 'women' because those were mostly used in the sentences/data for training and detecting misogynistic tweets/comments. Therefore, the models are biased towards such words and even normal tweets or comments containing such words are predicted as misogynistic. Similarly, Alves et al. [21] worked on highlighting unintended bias in tabular data of credit risk prediction (e.g., should not predict people from a specific ethnicity or region) or data about law students passing or failing a bar exam (e.g., race, sex, and family income).

The majority of the above, are introduced in the data-gathering process. However, we aim to highlight and discuss in detail the detection of the different types of bias that already exist in our datasets, specifically, in features of textual format. This typically entails data provided by customers, examples of which are: call transcripts, messages/emails, and survey responses, among others. Such data sources are unstructured (or semi-structured), totally depend on the input of customers, and are susceptible to all kinds of linguistic biases that are not captured in the list above.

B. AVAILABLE DATASETS

For textual elements of customer interaction-related datasets (e.g., reviews, feedback), the focus is on identifying gender, racial, religious, and client/consumer/customer bias. To this end, we have collected a few datasets to help us benchmark different types of bias detection algorithms. [22] reports findings in societal bias, more concisely regarding gender, race and religion. Gender bias was explored by looking at associations between gender and occupation; racial bias was explored by looking at how race impacted sentiment; religious bias by looking at which words occurred together with religious terms related to certain religions.

1) GENDER BIAS DATASETS

This type of bias is one of the most common and we have identified three datasets for our experiments. Hateful Symbols [23] dataset contains tweet IDs of tweets that are potentially sexist or racist. Actual tweets can then easily be retrieved using any Twitter API or parsing tweet content from https://twitter.com/anyuser/status/tweetId. Sexist compliment [24] includes tweet IDs of some tweets that contain hostile versus benevolent sexism. Misogyny identification[2] includes raw tweets and classifies them as misogynistic and/or aggressive [20].

2) RACIAL AND RELIGIOUS BIAS DATASETS

US hate crime[3] includes race/ethnicity and religious information on victims of hate crimes in the USA, as well as the various offence types. COMPAS[4] [25] (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendants' likelihood of re-offending (recidivism). It has been shown that the algorithm is biased in favour of white defendants, and against black inmates. The COMPAS dataset contains outcomes within 2 years of the decision, for over 10,000 criminal defendants in Broward County, Florida. The "Stop, Question and Frisk" database[5] contains data from NYPD officers' interactions with potential suspects of committing a crime. Features include locality-based information like time, street name, area code, etc.; crime-related features include weapons carried, contraband found, summons issued, suspect frisked, etc. The data also contains elements describing the physical appearance of the suspect like height, weight, build, hair colour, age, and race. The target variable of interest is whether an arrest was made or not.

3) CLIENT/CUSTOMER/CONSUMER BIAS DATASETS

There are several available datasets containing customer reviews of various products in textual form. A large number of those comes from Amazon reviews,[6] that besides a numerical rating, can provide a short description of the customers' experience with the product. Different types of bias may be present in such datasets, especially in cases where customer experience with it, was unsatisfactory. Another source of textual data comes from transcripts taken from calls between customers and call centre agents. The Action-Based Conversations Dataset [26] contains human-to-human interactions with 55 distinct user intents requiring unique sequences of actions. Although the actions are unrelated to our interests, detecting bias in those conversations is totally possible (and necessary). The development of reliable chatbot services has

[2]https://amievalita2020.github.io/data/
[3]https://www.kaggle.com/datasets/sumaiaparveenshupti/us-hate-crime-dataset-20102019-multiple-sources
[4]https://www.kaggle.com/datasets/danofer/compass
[5]https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
[6]https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products

been a major focus for many companies. To achieve that, they use chat datasets containing chat conversations between customers and call centre agents. Such instances include the Ubuntu Dialogue Corpus[7] containing about a million conversations on Ubuntu-related technical support issues, as well as the Twitter Customer Support dataset[8] containing over 3 million tweets and respective replies from some of the biggest companies like Amazon, Spotify and British Airways.

IV. REVIEW OF THE TOOLS/LIBRARIES

The use cases outlined in Section II involve generating systems that can analyse customer service transcripts by classifying biased language, performing sentiment analysis in customer dialogues, and recommending or selecting options based on the prediction and classification of a result as desirable or undesirable. In the interest of building high-level language models that are either unbiased themselves or capable of classifying bias and sentiment, it is necessary to examine both the steps to construct a language model as well as the biases that exist and are compounded throughout the process. High-level language models may inherit bias from their building corpora or the embeddings that allow them to process text, or from the training data for specific tasks.

There are several steps for constructing language models capable of performing tasks for the use cases of sentiment analysis, hate speech classification, and mitigating bias in prediction and recommender systems. To perform these sorts of tasks, a model must first be able to parse raw text into tokens or structures able to be understood by a computer, that is, sentence embedding or encoding. Sentences may be encoded using language models at the contextual embedding level (e.g., ELMo, CoVe), or at higher levels (e.g., BERT trained on additional task corpora). Contextual embedders, in turn, may be built from static embeddings, which encode words instead of sentences, or from text corpora; static embeddings themselves are built on text corpora. Higher-level language models may be able to learn additional tasks by training on datasets for, as an example, coreference resolution, with these task-training datasets understood as task-specific corpora.

Unfortunately, social biases (e.g., racism, sexism, xenophobia, ageism) exist in society and are reflected in the stories our society tells and has told throughout history. These stories and articles and comments that reflect our society's biases, however, make up much of the text corpora used to build and train language models, such that the language models at any step in the workflow inherit those social biases when encoding information, revealing themselves in disparate treatment and impact when that information is used by the model to make decisions and create outputs later down the line.

There have been biases found in all stages of the workflow, from the building corpora to the embeddings to the task corpora to the models performing the tasks. In addition, there have been flaws found even in the different bias detection and debiasing methods, which also must be acknowledged.

A. STATIC EMBEDDERS

Static embedders train on text corpora to create word embeddings, that is, vector representations of words based on what words occur around them in a corpus. Most frequently, these encodings are based on proximity and thus are vectors of association. Word2vec [27], one such embedder, creates embeddings using context windows around target words in corpora, and GloVe [28], another, uses the global co-occurrence statistics of words in a corpus. One notable exception to this trend, however, is the Symmetric Pattern [29] embedder, whose embedding construction relies on finding patterns of synonyms and antonyms in corpora.

Word2vec is one of the most widely used static embedders, with both the continuous-bag-of-words (CBOW) model and the skip-gram model as options to train the continuous word vectors from corpora, then an n-gram neural network language model (NNLM) is trained on the word representations. The CBOW uses the context of a word to encode it and predicts the word based on its context. The skip-gram model also encodes a word using its context, but weights the words by proximity, and predicts surrounding words based on the given word. These vectors are trained on the Google News corpus.

Global Vectors (GloVe) is also a commonly used embedder, which relies on statistics of global word-word co-occurrence counts in a corpus. Given a context window and a choice of left or right context, a matrix of co-occurrence counts is constructed. Different versions available are trained on the corpora of the 2010 Wikipedia dump, the 2014 Wikipedia dump, Gigaword 5th Edition, a combination of Gigaword 5th Edition + Wikipedia 2014, or the CommonCrawl.[9] The most popular versions of GloVe used in subsequent studies tend to be the Wikipedia+Gigaword combination, as well as the CommonCrawl version. Gender Neutral GloVe (GN-GloVe) [30], trained on the 2017 English Wikipedia dump, learns word embeddings with protected attributes by simultaneously identifying gender-neutral words and building the word vectors.

FastText [31] is based on the aforementioned skip-gram model, where each word is represented as a bag of character n-grams. A vector representation is associated to each character n-gram, and words are expressed as sums of these representations. This embedder is trained on Wikipedia dumps, and due to its sub-word encoding nature, it can compute word representations for words that did not appear in its training data.

Symmetric Pattern addresses the issue that word vector space representations are commonly more association-based than similarity-based; it instead looks for structures of synonyms and antonyms to extract meaning rather than mere

[7]https://www.kaggle.com/datasets/rtatman/ubuntu-dialogue-corpus

[8]https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter

[9]https://commoncrawl.org/

proximity. The training corpus is an 8G word corpus, including a 2012+2013 news article crawl [32], the One Billion (1B) Word Benchmark [33], the UMBC corpus [34], and the 2014 English Wikipedia dump. There are other static embedding language models, but these are the most relevant for both popularity and our use case.

B. CONTEXTUAL EMBEDDERS

Contextual embedders train on either text corpora or on existing static embeddings, usually those produced by one of the versions of GloVe. Instead of focusing merely on the statistical co-occurrences of words in corpora, contextual embedders use the context of words as well. Some contextual embedders, due to the nature of their construction, allow for the addition of further layers to turn the embedder into a higher-level language model capable of NLP tasks.

Contextual Vectors (CoVe) [35] is an encoder based on a two-layer, bidirectional long short-term memory (LSTM) network built off of GloVe vectors. InferSent [36] encoder is based on a bidirectional LSTM architecture with max pooling, also built on GloVe vectors trained on CommonCrawl, and further trained on the Stanford Natural Language Inference (SNLI) dataset [37].

The Universal Sentence Encoder (USE) [38] is a deep averaging network (DAN) encoder built on Wikipedia, news, question-answer pages, and forums and further trained on the SNLI corpus. Embeddings from Language Models (ELMo) [39] is another two-layer bidirectional LSTM language model, trained on the 1B Word Benchmark dataset.

Bidirectional Encoder Representations from Transformers (BERT) [40] is a bidirectional transformer encoder, trained on masked language model and next sentence prediction, trained on BookCorpus and English Wikipedia dumps. RoBERTa [41] increased and extended BERT's training time and data to include CommonCrawl News,[10] Stories [42], and OpenWebText [43], modified BERT's masking pattern, and removed the Next Sentence Prediction task.

Relevant models from the Generative Pre-Training (GPT) family include GPT [44], a transformer decoder trained on a unidirectional language model built off BookCorpus, and GPT-2 [45], a transformer decoder trained on a unidirectional language model built off WebText.

There are many other contextual embedding models, including those built using BERT, but the above are most relevant from popularity and for our use cases. It is important to highlight that these embedders, though often built on Wikipedia or CommonCrawl, can be trained on a vast variety of text corpora.

C. BIAS IN CONSTRUCTION CORPORA

Before describing the biases found within embeddings as well as how to detect them, highlighting studies which found bias in the corpora used to build and train these embedders can provide insight into the origin of some of the embedding biases. Some of these corpora biases are evaluated by examining a corpus itself, whereas other methods train an embedder on different corpora and compare differences in the performance of the final embeddings.

The work in Zhao et al [46] found that the 1B Word Benchmark used to train Symmetric Pattern as well as ELMo had a severely skewed representation of gender which could contribute to gender and gender-occupation bias. There were triple the occurrences of masculine pronouns than feminine pronouns, and those masculine pronouns would occur with occupation words more frequently than feminine pronouns would, regardless of whether the occupation had a gender stereotype in a certain direction. The authors in Tan and Celis [47] also evaluated the 1B Word Benchmark used to train Symmetric Pattern and ELMo, WebText used to train GPT-2, BookCorpus as used to train GPT and BERT, and a Wikipedia dump used to train Symmetric Pattern, FastText, GN-GloVe, and some iterations of GloVe. In line with [46], they found that feminine pronouns were even less common than gender-neutral or collective pronouns.

Unlike the preceding studies which evaluated the corpora directly, Chaloner and Maldonado [48] trained three different iterations of skip-gram embeddings on each of GoogleNews [27], Twitter posts [49], and PubMed Central Open Access subset (PMC) [50], as well as trained FastText embeddings on the Wikipedia GAP corpus [51]. Diaz et al. [52] explored age-related bias in sentiment analysis and examined GloVe trained on Wikipedia 2014 and Gigaword 5th Edition, trained on CommonCrawl, and trained on Twitter tweets. They found Twitter embeddings to produce the greatest bias, followed by CommonCrawl, with Wikipedia producing the least. Reference [53] used OpenAI WebText and OpenWebText Corpus (OpenAI-WT and OWTC) as used to train GPT-2 and RoBERTa to construct their RealToxicityPrompts dataset paired with Perspective API's[11] toxicity scores for the content. Reference [54] found that the SNLI dataset used to train InferSent and USE contained gender, age, race/ethnicity, and nationality bias.

The Colossal Clean Crawled Corpus (C4) [55], though not included in the above examples, was used to train GPT-3 [22] and investigated in Dodge et al. [56]. The nature of the cleaning algorithm's censorship indicates trends that are likely applicable to other cleaned Web corpora as well. To investigate the nature of the excluded content, the study clustered a random sample of excluded documents using the k-means algorithm and found that a little less than a third of the excluded documents were of a sexual or inappropriate nature, while the rest were on the topics of science, health, medicine, law, and politics. The study found that the cleaning algorithm was more likely to exclude documents from Black and Hispanic authors and documents mentioning sexual orientations than other demographics.

[10] https://commoncrawl.org/2016/10/news-dataset-available/

[11] https://perspectiveapi.com/

D. EVALUATING BIAS IN EMBEDDINGS

Word embeddings have been shown to encode and reflect the biases of the human corpora they are trained on. There are several methods that have been used to evaluate bias in static embeddings, with analogy completion and WEAT [57], as well as their derivatives, being the most popular. Additionally, Gonen and Goldberg [58] proposed widely used bias evaluation methods of quantifying the clustering accuracy of biased words, as well as evaluating the k-nearest neighbours of occupation words to detect gender bias.

The work in Mikolov et al. [27] tested the word2vec embeddings for semantic and syntactic soundness using analogy tasks including those generated by the 3COSADD function; Bolukbasi [14] evaluated embedding biases based on crowd-sourced judgments of analogies generated from the embeddings. The principle behind using analogies is that certain semantic and syntactic properties that are encoded into the word embeddings can be revealed through linear transformations and combinations of the word vectors.

The Word-Embedding Association Test (WEAT) [57] computes the cosine similarity score of encoded word vectors as a metric for the correlation between sets of target words, such as those that reflect demographic identities, and attribute words, which reveal sentiments associated with the identities. The Word-Embedding Factual Association Test (WEFAT) evaluates these associations against real-world statistics of, for example, gender representation in different occupations. The Unsupervised Bias Enumeration (UBE) [59] method. To reduce the reliance on the original WEAT's word lists for assessing bias, this approach defines groups by clustering normalized word vectors of names, defines word categories by clustering the most frequent word embedding tokens, selects words for the tests for each group of names, and finally computes p values for the associations found and ordering the tests based on the scores.

The Embedding Coherence Test (ECT) [60] evaluates how k-nearest neighbours change for gendered words and words with stereotypical gender associations upon normalisation of the embeddings, and the Embedding Quality Test (EQT) quantifies how bias in analogies improves upon implementation of a debiasing strategy. The Sentence Inference Retention Test (SIRT) [61] uses NLI principles to evaluate the mitigation of bias in embeddings as well as how well those debiased embeddings retain important information, such as the ability to correctly match the test's labels of entailment and contradiction.

The authors in Bolukbasi [14] proposed the concept of a gender subspace for vectors that can be used to assess bias, namely in the difference in Euclidean distances of occupation word vectors to male words as opposed to female counterparts. Many studies have used or built off of this concept of the gender or bias subspace. Additionally, the study proposes a calculation for Direct Bias, the proximity of a set of words to a gender vector. The work in Manzini et al. [62] expands the principle of the bias subspace to a multi-class setting by taking a ''one versus rest'' classifier approach instead of a linear word vector separability approach and introduces mean average cosine similarity (MAC) to evaluate bias in word sets.

CEAT [63] evaluated both static and contextual embeddings for intersectional biases, as well as proposed the Contextualized Embedding Association Test (CEAT) to specifically evaluate contextual embeddings. CEAT uses a random effects model and uses Combined Effect Size CES for its metric. The study also proposed methods for Intersectional Bias Detection (IBD), identifying words associated with intersectional identities, and Emergent Intersectional Bias Detection (EIBD), identifying those emergent intersectional biases in embeddings. After first using the methods on static embeddings to identify keywords, so as not to rely on attribute lists like WEAT does, those identified words can then be used to measure the biases in contextual embeddings, including use in CEAT.

E. DOCUMENTED BIAS IN EMBEDDINGS

Using the above, as well as other methods, several studies have identified different kinds of social biases in static word embeddings. Bias across embeddings can vary due to the corpora used to train the embeddings, or due to the nature of the embedder's strategy for extracting word vectors from corpora.

Several studies have found racial/gender bias in GloVe, trained on various corpora like CommonCrawl, Wikipedia2014, Gigaword5 [47], [57], [59], [63], [64], [65]. The work in Dev and Phillips [60], evaluated GloVe trained on Wikipedia and found names to encode not only gender or ethnicity but age as well, with bias affecting negative sentiment instead of occupational association. Other works also evaluate word2vec embeddings [10], [59], [62], [66]. The authors in Rozado [67] in particular, found negative associations with ''old age, middle or working-class socioeconomic status and below average physical appearance.''

ELMo, BERT and GPT have also been found to contain racial, gender and occupational bias by a number of studies [46], [47], [68], [69], [70], [71]. Examining social and intersectional biases, study in Guo and Caliskan [63] found ELMo to be the most biased contextual embedder, followed by BERT, GPT, and GPT-2, which corresponds to each model's level of contextualization. Different bias evaluation methods can reveal different kinds of biases; this is important to keep in mind when considering both bias evaluation as well as debiasing methods, which must be evaluated for bias themselves.

V. COMPARISON OF DIFFERENT APPROACHES FOR THE CONCERNED DATASETS/USE CASES

A. PERFORMANCE METRICS FOR BIAS DETECTION

Not all tools and solutions we discussed throughout our literature are fit for the diverse bias detection use cases that we have come across. Every tool has some advantages over others and may fit better in some scenarios. In this section, we present our findings from the comparison of the discussed

tools based on multiple aspects such as the provision of debiasing strategies, explainability features, multiple types of fairness metrics, etc.

Table 1 lists some of the different metrics used in algorithms throughout the literature for bias detection. Most works/tools also embed some type of disparity metric which is the ratio of a metric for a hypothesised privileged group over the respective metric of the unprivileged group. As an example, which model to use that results in more false positives or more false negatives [73], [74] as it will affect respective classes of people. This comes more under the Fair AI concepts. Another example could be cosine similarity (WEAT) is considered for each target (e.g., occupation) with respect to its attributes (e.g., gender).

B. COMPARISONS

Here, we compare the performance of different bias detection tools on customer service use case datasets. In specific, we use the following text-based datasets that; we believe, cover the most typical sources of textual data that can be utilised by a contact centre.

- Customer reviews.[12]
- The Action-Based Conversation Dataset (ABCD) [26].
- The Twitter customer support dataset.[13]

We further categorise the tools used based on the analysis method, and in particular, whether they use a sentiment analysis or association test approach. This way comparisons are more direct and fair for the different datasets, which in turn are sampled, in order to have a roughly similar sample count of around 10K data points.

1) SENTIMENT ANALYSIS

The tools used to evaluate sentiment on our selected datasets are the following:

- Language Interpretability Tool (LIT)[14]
- The Siebert model for sentiment analysis [76].

A major advantage is that both tools make sentiment predictions based on soft probabilities, making it easy to map each prediction on a scale from 0 (most negative sentiment score) to 1 (most positive sentiment score). Our aim is to demonstrate the differences (if any) between the tools, in order to highlight the existence of varying sentiments that are present in the selected datasets. Note that mitigating the existence of highly negative and possibly biased samples in the datasets, is out of the scope of this comparison.

The sentiment score distributions are shown in Figures 2a, 2b and 2c, one for each respective dataset. Looking at the predictions of the Siebert model, they are mostly close to the extremes for the reviews and ABCD datasets, while for the tweets, it gets more difficult to discern sentiment, despite the fact that the Twitter channels are mostly used to express complaints for different products. On the other hand, LIT appears to have a more realistic result overall, with the scores being more evenly distributed across the board. Interestingly, for the ABCD dataset, most conversations are mainly transactional, i.e., a more neutral result is anticipated, but that is not shown for either solution.

2) ASSOCIATION TESTS

For the association tests, we are comparing the following models.

- A WEAT-based model [57], trained on data containing either gender or toxic language bias.
- Dbias [77], a general model for bias detection, that outputs the bias-generating words in the text.
- A bias specific version[15] of the RoBERTa model [78].

The results displayed in Figure.3 show inconsistencies in the bias detection capabilities of the various models and datasets. Whether those are due to the inherent difficulties of detecting sentiment and bias in text, or imperfections of the tools, it is apparent that the customer experience use cases can benefit from more fine-tuned solutions for detecting bias, catered to their intricate needs.

VI. CHALLENGES AND OPPORTUNITIES FOR RESEARCH

A. CHALLENGES

Based on the literature review it is clear that considering the exponential growth of data and its usage in ML models for various applications, also increases the challenges in handling that data properly to avoid or mitigate its consequences. In this section, we will talk about a few of the challenges that can be found in bias detection.

1) LIMITATION IN UNDERSTANDING AND DETECTION OF BIAS

Section III discusses various types of Bias which shows that there are several types of biases whose root causes are different. Either it can be completely unintended and might result in data collection, storage or processing and development of an ML model stage or it can be intentional (both for unfair benefits or unfair enforcement rule of law which gives exemption to various rules or actions.) In some cases, the lack of a dataset for analysis or un-acceptance of the probability that there might be bias in an organisational increases the difficulty in detection, mitigation, or avoidance of bias in data and their respective models.

2) LIMITATION IN AVOIDANCE & MITIGATION OF BIAS

Since the literature agrees on the existence of Bias in various forms and at various stages, the research is lacking behind for its avoidance and mitigation. Currently, there are several works highlighted for individual cases that work at certain stages, work at data collection and pre-processing stage to

[12] https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products

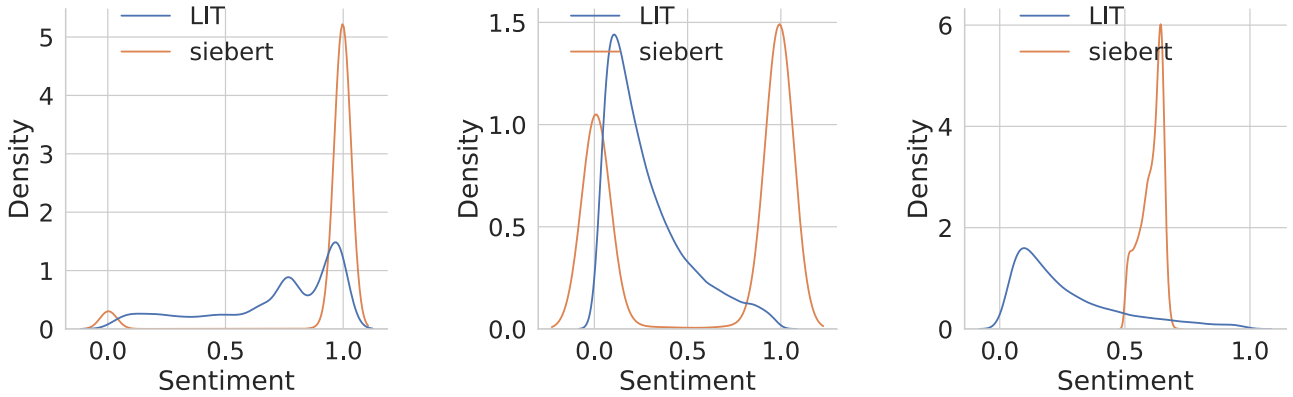
[13] https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter

[14] https://pair-code.github.io/lit/tutorials/tcav/

[15] https://huggingface.co/distilroberta-base

TABLE 1. Metrics used in bias detection algorithms.

Metric	Description	Citations
Cosine similarity (WEAT)	The sum of the Cosine Similarity of each target (e.g., occupations) with respect to its attributes (genders)	Caliskan et al. [57]
Relative norm distance	It captures the relative strength of association of a set of neutral words with respect to two groups,	Garg et al. [10]
Language Modeling Score	It is defined as the percentage of instances in which a language model prefers the meaningful over the meaningless association. The meaningful association corresponds to either the stereotype or the anti-stereotype option.	Nadeem et al. [72]
Stereotype Score	The percentage of instances in which a model prefers a stereotypical association over an anti-stereotypical association.	Nadeem et al. [72]
Idealised CAT score	Model comparison metric balancing between lms and ss.	Nadeem et al. [72]
False Positive/Negative Rate	The fraction of false positives/negatives of a group within the labelled negatives/positives of the group.	Saleiro et al. [73], Bellamy et al. [74]
False Discovery/Omission Rate	The fraction of false positives/negatives of a group within the predicted positive of the group.	Saleiro et al. [73], Bellamy et al. [74]
4/5th test	The pass rate of the lowest-population group has to be within 4/5ths of the pass rate of the highest population group	Pymetrics [75]



(a) Amazon reviews (b) ABCD (c) Twitter customer support

FIGURE 2. Probability density comparison for LIT and siebert models, for different datasets.

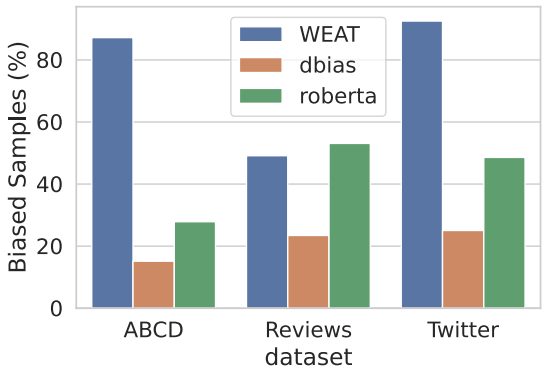


FIGURE 3. Percentage of biased samples for each dataset and bias detection algorithm.

avoid Bias, other works at mitigation of Bias while model training and works can be used for detection of bias after a model is trained, works at defining certain bias specific metrics to make the model fairer [79], [80], [81], [82], [83]. As complete removal of Bias is not possible, work is needed that can be a generic approach to link all the stages of ML pipeline which can avoid and mitigate the existence of bias.

B. OPPORTUNITIES

Considering the call centre and customer review use case, the following are a few potential opportunities to avoid and detect bias:

1) SCHEMA OR A MODEL TO MITIGATE BIAS

In industry, the use of model cards and AI OPs is increasing. The paper highlighted that bias exists at each level of the ML pipeline, hence a generic machine-readable vocabulary for model cards for AI Ops can be proposed. This schema can be adopted for detecting and mitigating bias throughout the pipeline. The aim can be to help investigate issues like unfair bias at data collection, pre-processing, algorithm selection, and performance metric selection in order to answer whether a model will perform consistently across a diverse range of people, or does it vary in unintended ways as characteristics like race or sex? Such a schema can bring clarity to these kinds of disparities, encouraging developers to consider their impact on a diverse range of people from the start of their planning to the development process, while keeping them in mind throughout the performance metric selection.

2) BIAS DETECTION USING SENTIMENT ANALYSIS

The use case II-A discusses data in the form of surveys and tabular data that can be uploaded by customers of a call centre and are not explored for bias detection. Such data can be used to detect bias using an algorithm designed for sentiment analysis [84]. Hence sentiment analysis is done at sentence [85], paragraph [86], and whole document level [87], and the same can be explored in this scenario.

3) BIAS DETECTION USING EMOTION RECOGNITION

The second use case II-B highlights the agents' bias towards the customer. Such bias can be detected by understanding the emotion of the agents' voices, chats, or emails. As an example, different authors Han et al. [88], [89] extracted salient features (e.g., expression of negative emotions, justifications, threats, from the emails of customers to detect their emotions. Similarly, Galanis et al. [90], Park et al. [91] shows that emotions from the voice recordings can be used to detect a bias towards certain customers. And a pattern of such detections can be used as a confirmation of bias in agents towards certain customers.

VII. CONCLUSION

This study introduced the existing landscape for detecting bias in data and also examined a number of different approaches that attempt to quantify and explain the biases within. It also focused on specific use cases that would be applicable for use with the tools and techniques which were researched and also the types of bias that can be found within these use cases. We concluded that not all tools and techniques highlighted and discussed are not fully applicable to the defined use cases we established early on with a subset of these having a better fit for some scenarios over others. Some potential challenges and future directions were also given that will be helpful for future research in this area.

ACKNOWLEDGMENT

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement Nos SFI/12/RC/2289_P2 and 20/SP/8955 at the Insight SFI Research Centre at the University of Galway. Insight, the SFI Research Centre for Data Analytics is funded by Science Foundation Ireland through the SFI Research Centres Programme. *(Andy Donald, Apostolos Galanopoulos, Edward Curry, Emir Muñoz, Ihsan Ullah, M. A. Waskow, Maciej Dabrowski, and Manan Kalra contributed equally to this work.)*

REFERENCES

[1] N. Sutaria, "Bias and ethical concerns in machine learning," *ISACA J.*, vol. 4, pp. 1–4, Jan. 2022.
[2] F. K. Došilovic, M. Brčić, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 210–215.
[3] (2020). *Cambridge Dictionary*. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/bias
[4] D. Levitin, *A Field Guide to Lies and Statistics: A Neuroscientist on How to Make Sense of a Complex World*. London, U.K.: Penguin, 2016.
[5] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," 2022, *arXiv:2202.00126*.
[6] H. Sun, G. Xu, J. Deng, J. Cheng, C. Zheng, H. Zhou, N. Peng, X. Zhu, and M. Huang, "On the safety of conversational models: Taxonomy, dataset, and benchmark," 2021, *arXiv:2110.08466*.
[7] M. Kim, O. Reingold, and G. Rothblum, "Fairness through computationally-bounded awareness," in *Proc. Adv. NeurIPS*, vol. 31, 2018, pp. 1–12.
[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, Jan. 2012, pp. 214–226.
[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.
[10] 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. E3635–E3644, Apr. 2018, doi: 10.1073/pnas.1720347115.
[11] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1841–1844.
[12] *Why Call Centers Are Critical to Digital Experience Platforms*, OpenText, Waterloo, ON, Canada, 2021.
[13] V. Chernozhukov, C. Cinelli, W. K. Newey, A. Sharma, and V. Syrgkanis. (2021). *Omitted Variable Bias in Machine Learned Causal Models*. [Online]. Available: http://hdl.handle.net/10419/260381
[14] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 4356–4364.
[15] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: ACM, 2021, pp. 1–9.
[16] W. Jeong, K. Lee, D. Yoo, D. Lee, and S. Han, "Toward reliable and transferable machine learning potentials: Uniform training by overcoming sampling bias," *J. Phys. Chem. C*, vol. 122, no. 39, pp. 22790–22795, Oct. 2018, doi: 10.1021/acs.jpcc.8b08063.
[17] B. French and P. J. Heagerty, "Analysis of longitudinal data to evaluate a policy change," *Statist. Med.*, vol. 27, no. 24, pp. 5005–5025, Oct. 2008, doi: 10.1002/sim.3340.
[18] S. Barbosa, D. Cosley, A. Sharma, and R. M. Cesar, "Averaging gone wrong: Using time-aware analyses to better understand behavior," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, Art. no. 829841, doi: 10.1145/2872427.2883083.
[19] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *SSRN Electron. J.*, vol. 5, p. 13, Jul. 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2019.00013
[20] D. Nozza, C. Volpetti, and E. Fersini, "Unintended bias in misogyny detection," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Oct. 2019, pp. 149–155.
[21] G. Alves, M. Amblard, F. Bernier, M. Couceiro, and A. Napoli, "Reducing unintended bias of ML models on tabular and textual data," in *Proc. IEEE 8th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2021, pp. 1–10.
[22] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[23] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93. [Online]. Available: http://www.aclweb.org/anthology/N16-2013
[24] A. Jha and R. Mamidi, "When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data," in *Proc. 2nd Workshop NLP Comput. Social Sci.*, 2017, pp. 7–16.
[25] T. Brennan and W. Dieterich, "Correctional offender management profiles for alternative sanctions (COMPAS)," in *Handbook of Recidivism Risk/Needs Assessment Tools*. Hoboken, NJ, USA: Wiley, 2017, pp. 49–75.
[26] D. Chen, H. Chen, Y. Yang, A. Lin, and Z. Yu, "Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 3002–3017. [Online]. Available: https://www.aclweb.org/anthology/2021.naacl-main.239
[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[28] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[29] R. Schwartz, R. Reichart, and A. Rappoport, "Symmetric pattern based word embeddings for improved word similarity prediction," in *Proc. 19th Conf. Comput. Natural Lang. Learn.*, 2015, pp. 258–267. [Online]. Available: https://aclanthology.org/K15-1026

[30] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4847–4853. [Online]. Available: https://aclanthology.org/D18-1521

[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[32] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, and L. Specia. (Jun. 2014). *Proceedings of the Ninth Workshop on Statistical Machine Translation*. [Online]. Available: https://aclanthology.org/W14-3300

[33] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," 2013, *arXiv:1312.3005*.

[34] T. F. J. M. L. Han, L. A. Kashyap, and J. Weese, "UMBC-CORE: Semantic textual similarity systems," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics*, Jun. 2013, pp. 1–22.

[35] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," 2017, *arXiv:1708.00107*.

[36] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. EMNLP*, Sep. 2017, pp. 670–680. [Online]. Available: https://aclanthology.org/D17-1070

[37] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 632–642.

[38] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," 2018, *arXiv:1803.11175*.

[39] M. Peters, "Deep contextualized word representations," in *Proc. NAACL-HLT*, 2018, pp. 1–12.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[41] 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*.

[42] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," 2018, *arXiv:1806.02847*.

[43] A. Gokaslan and V. Cohen. (2019). *Openwebtext Corpus*. [Online]. Available: http://Skylion007.github.io/OpenWebTextCorpus

[44] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018. [Online]. Available: https://www.bibsonomy.org/bibtex/273ced32c0d4588eb95b6986dc2c8147c/jonaskaiser

[45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, San Francisco, CA, USA, 2018. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[46] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2019, pp. 629–634. [Online]. Available: https://aclanthology.org/N19-1064

[47] Y. C. Tan and L. E. Celis, *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Red Hook, NY, USA: Curran Associates, 2019.

[48] K. Chaloner and A. Maldonado, "Measuring gender bias in word embeddings across domains and discovering new gender bias word categories," in *Proc. 1st Workshop Gender Bias Natural Lang. Process.*, 2019, pp. 25–32. [Online]. Available: https://aclanthology.org/W19-3804

[49] F. Godin, B. Vandersmissen, W. D. Neve, and R. V. de Walle, "Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for Twitter microposts using distributed word representations," in *Proc. NUT@IJCNLP*, 2015, pp. 146–153.

[50] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, "How to train good word embeddings for biomedical NLP," in *Proc. 15th Workshop Biomed. Natural Lang. Process.*, 2016, pp. 1–4.

[51] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the GAP: A balanced corpus of gendered ambiguous pronouns," 2018, *arXiv:1810.05201*.

[52] M. Diaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle, "Addressing age-related bias in sentiment analysis," in *Proc. CHI*, 2018, p. 114, doi: 10.1145/3173574.3173986.

[53] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Proc. Findings Assoc. Comput. Linguistics*, Nov. 2020, pp. 3356–3369. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.301

[54] R. Rudinger, C. May, and B. Van Durme, "Social bias in elicited natural language inferences," in *Proc. 1st ACL Workshop Ethics Natural Lang. Process.*, 2017, pp. 74–79. [Online]. Available: https://aclanthology.org/W17-1609

[55] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[56] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 1286–1305. [Online]. Available: https://aclanthology.org/2021.emnlp-main.98

[57] 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, doi: 10.1126/science.aal4230.

[58] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," 2019, *arXiv:1903.03862*.

[59] N. Swinger, M. De-Arteaga, N. T. Heffernan, M. D. M. Leiserson, and A. T. Kalai, "What are the biases in my word embedding?" in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2019, pp. 305–311, doi: 10.1145/3306618.3314270.

[60] S. Dev and J. M. Phillips, "Attenuating bias in word vectors," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1–15.

[61] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, "OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 5034–5050. [Online]. Available: https://aclanthology.org/2021.emnlp-main.411

[62] T. Manzini, L. Y. Chong, A. W. Black, and Y. Tsvetkov, "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. Linguistics, Hum. Lang. Technol.*, 2019, pp. 615–621. [Online]. Available: https://aclanthology.org/N19-1062

[63] 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, Art. no. 122133, doi: 10.1145/3461702.3462536.

[64] A. Sutton, T. Lansdall-Welfare, and N. Cristianini, "Biased embeddings from wild data: Measuring, understanding and removing," 2018, *arXiv:1806.06301*.

[65] O. Agarwal, F. Durupınar, N. I. Badler, and A. Nenkova, "Word embeddings (also) encode human personality stereotypes," in *Proc. 8th Joint Conf. Lexical Comput. Semantics*, 2019, pp. 205–211. [Online]. Available: https://aclanthology.org/S19-1023

[66] 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. [Online]. Available: https://aclanthology.org/P19-1162

[67] D. Rozado, "Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types," *PLoS ONE*, vol. 15, no. 4, pp. 1–26, Apr. 2020, doi: 10.1371/journal.pone.0231189.

[68] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," in *Proc. NAACL-HLT*, 2019, pp. 622–628.

[69] M. Bartl, M. Nissim, and A. Gatt, "Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias," in *Proc. 2nd Workshop Gender Bias Natural Lang. Process.*, Dec. 2020, pp. 1–16. [Online]. Available: https://aclanthology.org/2020.gebnlp-1.1

[70] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 15–20. [Online]. Available: https://aclanthology.org/N18-2003
[71] C. Basta, M. R. Costa-jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," in *Proc. 1st Workshop Gender Bias Natural Lang. Process.*, 2019, pp. 33–39. [Online]. Available: https://aclanthology.org/W19-3805
[72] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5356–5371. [Online]. Available: https://aclanthology.org/2021.acl-long.416
[73] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," 2018, *arXiv:1811.05577*.
[74] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018, *arXiv:1810.01943*.
[75] Pymetrics. (2018). *Audit-AI*. Accessed: Jun. 23, 2021. [Online]. Available: https://github.com/pymetrics/audit-ai
[76] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Marketing*, vol. 40, no. 1, pp. 75–87, Mar. 2023.
[77] S. Raza, D. J. Reji, and C. Ding, "Dbias: Detecting biases and ensuring fairness in news articles," *Int. J. Data Sci. Anal.*, vol. 2022, pp. 1–21, Sep. 2022.
[78] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[79] S. Alelyani, "Detection and evaluation of machine learning bias," *Appl. Sci.*, vol. 11, no. 14, pp. 1–10, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/14/6271
[80] C. J. Pannucci and E. G. Wilkins, "Identifying and avoiding bias in research," *Plastic Reconstructive Surg.*, vol. 126, no. 2, pp. 619–625, Aug. 2010.
[81] K. Zhang, B. Khosravi, S. Vahdati, S. Faghani, F. Nugen, S. M. Rassoulinejad-Mousavi, M. Moassefi, J. M. M. Jagtap, Y. Singh, P. Rouzrokh, and B. J. Erickson, "Mitigating bias in radiology machine learning: 2. model development," *Radiol., Artif. Intell.*, vol. 4, no. 5, Sep. 2022, Art. no. e220010.
[82] Y. Zhao, P. Yin, Y. Li, X. He, J. Du, C. Tao, and Yi Guo, "Data and model biases in social media analyses: A case study of COVID-19 tweets," in *Proc. AMIA Annu. Symp.*, 2021, p. 1264.
[83] C. Pagel and C. A. Yates, "Tackling the pandemic with (biased) data," *Science*, vol. 374, no. 6566, pp. 403–404, Oct. 2021.
[84] A. H. Sweidan, N. El-Bendary, and H. Al-Feel, "Sentence-level aspect-based sentiment analysis for classifying adverse drug reactions (ADRs) using hybrid ontology-XLNet transfer learning," *IEEE Access*, vol. 9, pp. 90828–90846, 2021.
[85] Z. Lei, Y. Yang, and M. Yang, "SAAN: A sentiment-aware attention network for sentiment analysis," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 1197–1200.
[86] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–19, Dec. 2021.
[87] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, Jul. 2019. [Online]. Available: https://www.mdpi.com/2504-4990/1/3/48
[88] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6494–6498.
[89] N. Gupta, M. Gilbert, and G. D. Fabbrizio, "Emotion detection in email customer care," *Comput. Intell.*, vol. 29, no. 3, pp. 489–505, Aug. 2013.
[90] D. Galanis, S. Karabetsos, M. Koutsombogera, H. Papageorgiou, A. Esposito, and M.-T. Riviello, "Classification of emotional speech units in call centre interactions," in *Proc. IEEE 4th Int. Conf. Cognit. Infocommun. (CogInfoCom)*, Dec. 2013, pp. 403–406.
[91] K. Park, M. Cha, and E. Rhim, "Positivity bias in customer satisfaction ratings," in *Proc. Companion Web Conf. Web Conf.*, 2018, pp. 631–638.

ANDY DONALD is currently a Research Fellow and the Lead of the Applied Innovation Unit at Insight, University of Galway. His current research interests include the automation of the AI and machine learning process and the application of explainable AI methods within these processes and their application in real world, and cloud based situations. He has a particular interest in the development of surrogate models for complex computational models and how these can be applied to help real-time predictions. He has made significant contributions to research on distributed systems and stream processing and researching linked data and knowledge graph construction. His interests also extend into the management of large scale multimodal data ecosystems. His significant private sector background also provides a unique view into how innovative research can be scaled and applied to real world issues both at an academic and industry level.

APOSTOLOS GALANOPOULOS received the B.Sc. degree from the Department of Electrical and Computer Engineering, University of Thessaly, Volos, in 2014, the M.Sc. degree in science and technology of electrical and computer engineering from the University of Thessaly, in 2016, and the Ph.D. degree from the School of Computer Science and Statistics, Trinity College Dublin, Ireland, in 2021, focusing on resource management for data analytics over edge computing networks. In 2021, he joined Genesys, as a Machine Learning Research Engineer. His research interests include wireless communications, edge computing, optimization theory, and machine learning.

EDWARD CURRY is currently an Established Professor in data science and the Director of the Insight SFI Research Centre for Data Analytics, University of Galway. He has made substantial contributions to semantic technologies, incremental data management, event processing middleware, software engineering, and distributed systems and information systems. He combines strong theoretical results with high-impact practical applications. The excellence and impact of his research have been acknowledged by numerous awards, including best paper awards and the University of Galway President's Award for Societal Impact, in 2017. His team's technology enables intelligent systems for smart environments in collaboration with several industrial partners. He is an Organizer and the Program Co-Chair of major international conferences, including CIKM 2020, ECML 2018, IEEE Big Data Congress, and European Big Data Value Forum. He is the Co-Founder and the elected Vice President of the Big Data Value Association, an industry-led European big data community, has built consensus on a joint European big data research and innovation agenda, and influenced European data innovation policy to deliver on the agenda.

EMIR MUÑOZ received the B.Eng. and M.Sc. degrees in computer engineering from Universidad de Santiago, Chile, in 2009 and 2011, respectively, and the Ph.D. degree in knowledge graph mining from the National University of Ireland, Galway (now University of Galway), in 2020. He is currently a Senior Manager in machine learning with the Digital & AI Group, Genesys Cloud Services Inc., where he and his team research, develop, and innovate on AI/ML solutions for contact center optimization and customer experience. He serves as a reviewer for journals, conferences, and books, such as *Briefings in Bioinformatics*, *Semantic Web Journal*, *ACL*, and *DEXA*.

IHSAN ULLAH received the Ph.D. degree from the University of Milan, Italy, in 2017. Currently, he is an Assistant Professor with the School of Computer Science, University of Galway, Ireland. He is also a funded Investigator with the Insight Research Center for Data Analytics, Galway. His research interests include computer vision, deep learning, explainable AI, and social data.

M. A. WASKOW is a Research Assistant in the applied innovations unit at Insight, University of Galway. They completed their B.Sc. in biosystems analytics and technology at the University of Arizona, USA, with a concentration in Statistics and Data Science. Their research interests include systems analysis and modeling for ecological, climate, and agricultural contexts using machine learning strategies.

MACIEJ DABROWSKI is very passionate about building large-scale AI products that are easy to use and create value, he has been doing it both in research and industry for over ten years. Prior to joining Genesys, in 2018, he enjoyed the ups and downs of startup life driving AI development with Altocloud. He received education in engineering, business and decision support systems. He spends his free time scuba diving and travelling.

MANAN KALRA received the M.Sc. degree in data science and analytics from University College Cork, Ireland, in 2020. Currently, he is a Machine Learning Engineer and contributes to the predictive and automated routing capabilities of the contact center solutions offered by Genesys Cloud Services Inc. His research interests include AI ethics and computational social science.