



## Investigating transformer models for textual bias detection in model, data, and dataspace cards

Title	Investigating transformer models for textual bias detection in model, data, and dataspace cards
Author(s)	Donald, Andy; Galanopoulos, Apostolos; Kumar Ojha, Atul; Curry, Edward; Muñoz, Emir; Ullah, Ihsan; McCrae, John P.; Kalra, Manan; Saxena, Sagar; Iqbal, Talha
Publication Date	2026-01-28
Publisher	Springer
Repository DOI	<a href="https://doi.org/10.1007/s43681-025-00975-3">https://doi.org/10.1007/s43681-025-00975-3</a>



# Investigating transformer models for textual bias detection in model, data, and dataspace cards

Andy Donald<sup>1</sup> · Apostolos Galanopoulos<sup>2</sup> · Atul Kumar Ojha<sup>1</sup> · Edward Curry<sup>1</sup> · Emir Muñoz<sup>2</sup> · Ihsan Ullah<sup>1,3</sup> · John P. McCrae<sup>1</sup> · Manan Kalra<sup>2</sup> · Sagar Saxena<sup>2</sup> · Talha Iqbal<sup>1,3</sup>

Received: 4 September 2025 / Accepted: 26 December 2025  
© The Author(s) 2026

## Abstract

Identifying hidden biases in AI documentation metadata (model, data, and dataspace cards) is essential for responsible AI; yet this domain remains largely unexplored. The proposed work evaluates four Transformer models (XLNet, DistilBERT, RoBERTa, and ELECTRA) for bias detection across publicly available, synthetic, and custom datasets. On the BABE news corpus, all models achieved 77–80% accuracy, with only ELECTRA exceeding 80% on every metric. To address the absence of publicly available AI-card datasets, we generated synthetic metadata for two use cases (*Customer Interaction and Customer Data Uploaded by Organisations*) using ChatGPT. Models trained on this synthetic corpus displayed near-perfect scores, reflecting shared stylistic cues embedded in the generated text. To test real-world robustness, we curated a Hugging Face dataset by scraping documentation comments, filtering for bias-related keywords, and obtaining annotations from four independent labellers in a single-blind setting. Partial fine-tuning (zero-shot) evaluations of models trained only on BABE or synthetic data revealed substantial performance drops on this real-world set. To mitigate this cross-domain loss, we introduce a cascaded, full fine-tuning (few-shot) pipeline in which Transformer models are sequentially fine-tuned on BABE, synthetic text, and a subset of the Hugging Face corpus. Evaluation on the remaining portion achieved over 85% across all performance metrics, enhancing precision and generalisation. This study demonstrates the challenges of bias detection beyond controlled or synthetic data and highlights cascaded fine-tuning as a practical, low-resource strategy. Future directions include leveraging evidence fusion methods, integrating cross-attention with bias taxonomies, and adopting dual-encoder architectures to advance bias detection toward more in-depth, knowledge-guided reasoning.

**Keywords** Transformer models · Bias detection · Textual data · Model and data cards · Dataspace cards · Use cases

---

✉ Edward Curry  
edward.curry@universityofgalway.ie

✉ Ihsan Ullah  
ihsan.ullah@universityofgalway.ie

✉ Talha Iqbal  
talha.iqbal@universityofgalway.ie

Andy Donald  
andy.donald@universityofgalway.ie

Apostolos Galanopoulos  
Apostolos.Galanopoulos@genesys.com

Atul Kumar Ojha  
atul.ojha@universityofgalway.ie

Emir Muñoz  
emir.munoz@genesys.com

John P. McCrae  
john.mccrae@universityofgalway.ie

Manan Kalra  
manan.kalra@genesys.com

Sagar Saxena  
sagar.saxena@genesys.com

<sup>1</sup> Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Galway, Ireland

<sup>2</sup> Genesys Cloud Services Inc., Galway, Ireland

<sup>3</sup> School of Computer Science, University of Galway, Galway, Ireland

## 1 Introduction

The proliferation of Natural Language Processing (NLP) technologies across diverse domains highlights the necessity for robust, scalable methodologies to detect and mitigate linguistic biases in textual data. These biases, often implicitly embedded within language, can perpetuate stereotypes, misinformation, and discriminatory practices, thereby influencing public perception and decision-making processes [1]. The challenge is further compounded by the complexities inherent in human language, where context, nuance, and cultural factors play pivotal roles in shaping meaning [2].

Transformer-based language models have emerged as powerful tools in NLP, demonstrating remarkable capabilities in understanding and generating human-like text. Their proficiency in capturing contextual relationships makes them well-suited for tasks involving bias detection. However, the effectiveness of these models is intrinsically linked to the quality and diversity of the datasets on which they are trained [3]. These models remain susceptible to detecting and mitigating different types of biases embedded in their training data. This vulnerability is particularly concerning in the context of artificial intelligence (AI) documentation tools such as Model Cards, Data Cards, and Dataspace Cards, which aim to promote transparency and accountability in AI systems [4, 5].

Models such as RoBERTa, XLNet, ELECTRA, and DistilBERT have demonstrated varying degrees of effectiveness in identifying biases within textual data. The Word Embedding Association Test (WEAT) has been utilised to measure bias patterns across models like BERT, RoBERTa, and XLNet, revealing inherent biases in word associations [6]. Furthermore, previous studies [7–9] have examined the internal activations of transformer models to visualise and reduce hidden biases, leading to fairness-aware pruning techniques. These insights highlight the importance of selecting Transformer architectures that are not only expressive but also empirically robust across diverse downstream language tasks. In this context, models such as RoBERTa have demonstrated strong representational capacity in emotion recognition tasks on datasets like GoEmotion [10], while ELECTRA, whose replaced-token detection method allows efficient and stable pre-training as compared to traditional masked language modelling [11]. Transformer models provide suitable foundations for analysing how bias manifests and propagates in textual representations [12]. Their empirically established performance makes them well-suited for investigating bias detection in real-world and cross-domain contexts.

A significant limitation encountered in deploying transformer-based models for textual bias detection within these

Cards is the constrained availability of bias-annotated metadata relevant to datacards, model cards and dataspace cards. The scarcity of annotated comments affects the identification of nuanced biases and reduces the reliability of automated bias detection methods. Consequently, expanding and diversifying annotated datasets becomes critical to enhance transformer models' capabilities in identifying linguistic patterns indicative of biases, thereby strengthening the efficacy of documentation tools designed to foster fairness and transparency in AI systems.

Recent research has delved into adaptive prompting strategies to generate synthetic datasets and enhance bias detection capabilities [13–15]. For instance, an adaptive prompting approach predicts the optimal composition of discrete prompt techniques for each input instance, improving performance in social bias detection tasks [16]. Additionally, methods have been proposed to quantify and mitigate prompt bias, ensuring the reliability of benchmarks and enhancing the retrieval capabilities of prompts [17–19].

A significant gap exists between bias definitions and what different humans perceive as biased text. While synthetic data offers scalability and controlled conditions for model training, it often lacks the depth and variability inherent in human-generated content. This discrepancy can lead to models that perform well on synthetic benchmarks but do not perform well when tested with real-world data, where biases are more nuanced and context-dependent [20]. Bridging this gap requires aligning synthetic data generation processes with human language and cognition complexities.

Thus, the proposed study aims to investigate the efficacy of Transformer models (XLNet, ELECTRA, RoBERTa, DistilBERT) in detecting textual biases within the metadata in models, data, and dataspace cards. By addressing the limitations posed by dataset scarcity and the divergence between synthetic and human annotations, we aim to enhance the robustness and reliability of bias detection methodologies. Our approach focuses on creating a comprehensive dataset specifically for evaluating Transformer-based models. We began by evaluating Transformer-based models using a publicly available dataset, focused on biased news articles. However, since our primary objective was to assess the models' effectiveness in detecting bias within the model, data and dataspace cards (specifically in the contexts of *Customer Interaction* and *Customer Data Uploaded by Organisations* domain), we found no existing datasets aligned with these use cases. To address this gap, we created two custom datasets: one comprising synthetic examples and another based on real-world human annotations sourced from Hugging Face. This approach enabled a robust evaluation across both synthetic and real-world scenarios tailored to our specific use cases.

## 2 Methodology

This section describes the model architectures, highlights the key implementation procedures, and offers detailed summaries of the datasets employed. The figure illustrates the overall workflow for detecting hidden bias in AI documentation metadata (Fig. 1).

### 2.1 Model architectures

The Transformer-based models used in this study include XLNet, ELECTRA, RoBERTa, and DistilBERT. Each of these models offers distinct advantages for textual bias detection tasks. XLNet has permutation-based training, enabling comprehensive context modelling; ELECTRA leverages efficient pretraining techniques that promote faster convergence; RoBERTa benefits from optimised training strategies that enhance performance; and DistilBERT’s compact architecture supports deployment in resource-constrained environments. The following is a comprehensive description of each model:

#### 2.1.1 XLNet

XLNet, introduced by Yang et al. [21], is an autoregressive pre-trained model that integrates the strengths of autoencoder (BERT) and autoregressive (GPT) language models. Unlike BERT, which relies on masked language modelling, XLNet employs a permutation-based training objective that considers all possible permutations of the input sequence, enabling it to capture bi-directional contexts effectively [22]. This approach allows XLNet to model the joint probability of the sequence of tokens, enhancing its ability to understand complex language structures. Additionally, it incorporates the segment-level recurrence mechanism and relative positional encoding from Transformer-XL, enabling it to model longer-term dependencies in the text [23].

#### 2.1.2 ELECTRA

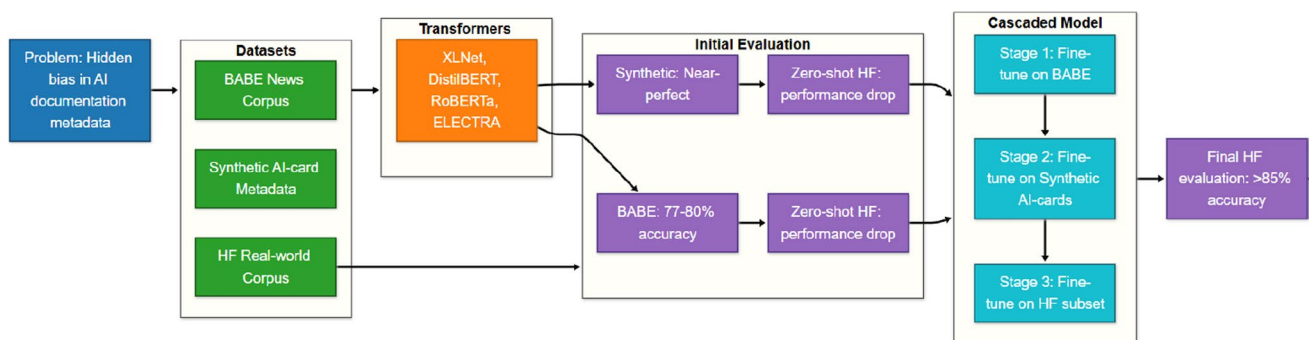
ELECTRA introduces a relatively novel pre-training method called replaced token detection, which differs from the traditional masked language modelling (BERT) approach [11]. ELECTRA’s architecture comprises two Transformer models: the generator and the discriminator, both utilising the standard Transformer encoder architecture. In this framework, a small generator model replaces some tokens in the input with plausible alternatives, and a larger discriminator model is trained to distinguish between the original and replaced tokens. This setup allows the model to learn from all input tokens rather than just the masked ones, resulting in more efficient training [11].

#### 2.1.3 RoBERTa

RoBERTa (Robustly Optimised BERT Approach) is an optimised variant of BERT that modifies key hyperparameters and training strategies to enhance performance. Notably, RoBERTa removes the next-sentence prediction objective, employs dynamic masking, and is trained on larger mini-batches with more data [24]. These changes enable RoBERTa to achieve better performance on various NLP tasks without altering the underlying Transformer encoder architecture.

#### 2.1.4 DistilBERT

It is a distilled version of the BERT model that aims to reduce the model size and increase efficiency while retaining most of BERT’s language understanding capabilities. Through knowledge distillation, the DistilBERT reduces the number of layers by 40%, resulting in a model that is 60% faster and 40% smaller than BERT, yet maintaining approximately 97% of the performance on the benchmark tasks [25]. This makes the DistilBERT model particularly suitable for applications with limited computational resources.



**Fig. 1** Workflow for detecting hidden bias in AI documentation metadata. The cascaded fine-tuning pipeline adapts transformer models from public and synthetic datasets to real-world metadata, supporting evaluation and FAIR AI implementation

## 2.2 Implementation details

The subsequent section describes the key implementation steps employed in designing the experiments conducted in this study.

### 2.2.1 Tokenisation

Each transformer model utilises a tokeniser technique that is tailored to its pertaining corpus and architectural design. Specifically, RoBERTa and DistilBERT utilise Byte-Pair Encoding (BPE), XLNet employs SentencePiece, and ELECTRA uses WordPiece tokenisation [26]. These tokenisers convert raw text into subword units, facilitating the handling of out-of-vocabulary (OOV) terms and enabling the models to process diverse linguistic inputs effectively. The Hugging Face Transformers library provides both standard and fast implementations of these tokenisers, with additional functionalities such as efficient mapping between original text and tokenised output [27].

### 2.2.2 Training pipeline

The implemented training pipeline for fine-tuning the transformer models on the bias detection task had the following key components. Although the underlying encoder/transformer model changes, all other components (data curation, optimisation strategy, evaluation, and artefact management) remain invariant, thereby facilitating fair and controlled comparisons across models.

**Data preprocessing** The input text (sentence/label pair) is converted into the tensor triplet (input\_ids, attention\_mask, and labels) expected by the transformers classification heads. The tokenisation is handled by the model-specific tokeniser (XLNet, ELECTRA, DistilBERT, or RoBERTa) with truncation and zero-padding to a fixed maximum length of 128 sub-word tokens, ensuring compatibility with the model architectures.

**Model configuration** For each experiment, a pre-trained checkpoint (*xlnet-base-cased*, *google/electra-base-discriminator*, *distilbert-base-uncased*, or *roberta-base*) is retrieved from the Hugging Face model repository. A lightweight linear classification head (output dimensionality = 2) is initialised and appended to the encoder to accommodate the binary classification task of bias detection.

**Training parameters** The training parameters were selected by hit-and-trial methods. All models were fine-tuned using the AdamW optimiser with learning rate =  $2 \times 10^{-5}$  and weight decay = 0.01. Each gradient step comprises: gradient buffer reset, forward pass, cross-entropy loss computation, back-propagation, and parameter update. Mean epoch-level loss is logged to monitor convergence

and to detect potential overfitting or learning-rate instability irrespective of the underlying model. The training epochs were set to 10 with a mini-batch size = 10 as well.

### 2.2.3 Evaluation metrics

To maintain class-balance neutrality, macro-averaged accuracy, precision, recall, and F1-scores are calculated.

In all the below equations, *CB* (Correct Bias) denotes the sentence *is* biased and the model predicts “biased”; *CN* (Correct Non-bias) denotes the sentence *is* non-bias and the model predicts “non-bias”; *FB* (False Bias) denotes the sentence *is* non-bias but the model predicts “biased”; *FN* (False Non-bias) denotes the sentence *is* biased but the model predicts “non-bias”; while  $R_n$  and  $P_n$  denotes recall and precision at the *n*-th threshold.

**Accuracy.** Measures the proportion of correctly predicted sentences over the total number of test sentences. Generally, accuracy can be misleading in imbalanced datasets.

$$\text{Acc} = \frac{\text{CB} + \text{CN}}{\text{CB} + \text{CN} + \text{FB} + \text{FN}} \quad (1)$$

**Precision.** Calculates the ratio of true positive predictions to total predicted positive examples, indicating the model’s ability to avoid false positives.

$$\text{Prec} = \frac{\text{CB}}{\text{CB} + \text{FB}} \quad (2)$$

**Recall (sensitivity).** Determines the ratio of true positive predictions to all actual positives, reflecting the model’s capacity to identify all relevant/biased sentences.

$$\text{Rec} = \frac{\text{CB}}{\text{CB} + \text{FN}} \quad (3)$$

**F1-Score.** Is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. This metric is particularly informative when dealing with class imbalance.

$$F1 = \frac{2 \text{CB}}{2 \text{CB} + \text{FB} + \text{FN}} \quad (4)$$

**AUC-PR (area under the precision-recall curve).** Quantifies the trade-off between precision and recall across the decision threshold. It is computed as the area under the precision-recall curve, which plots precision (the ratio of true positives to predicted positives) against recall (the ratio of true positives to actual positives). A higher AUC-PR

indicates better model performance, especially for imbalanced datasets where traditional ROC-AUC may be misleading. It is computed as the following equation, according to the scikit-learn library:

$$AP = \sum_n (R_n - R_{n-1})P_n \tag{5}$$

### 2.3 Data preparation

For the analysis, three distinct datasets were utilised: a publicly available dataset, a synthetically generated dataset, and a real-world comments dataset collected by scraping Hugging Face. The data preparation procedures for each dataset are detailed below:

#### 2.3.1 Neural media bias detection using distant supervision with BABE

The BABE dataset serves as a foundational resource for this study, offering high-quality annotations for media bias detection. BABE comprises 3,700 sentences extracted from U.S. news articles published between January 2017 and June 2020, covering a diverse range of topics and outlets [28]. Each sentence is annotated at both the word and sentence levels for bias, with annotations performed by experts with diverse backgrounds in computer science from different geographies to ensure consistency and reliability. This expert-driven approach addresses the limitations of previous datasets that relied on crowdsourcing, which often resulted in lower annotation quality and inter-annotator agreement [29] (Table 1).

To analyse annotation consistency and quality, the BABE dataset is divided into two sub-groups: Sub-group 1 had 8 annotators, while Sub-group 2 had 5 annotators. Sub-Group 1 consists of 1700 sentences annotated by eight experts, while Sub-Group 2 includes 3674 sentences annotated by five experts. The "No Agreement" category indicates instances where annotators could not reach a consensus on the bias label, highlighting the subjective nature of bias detection. The low number of disagreements in Sub-Group 2 suggests improved consensus among annotators, potentially

due to refined annotation guidelines or increased annotator experience.

#### 2.3.2 Synthetic sentence generation

Synthetic data generation has emerged as a viable approach to augment limited datasets. By leveraging large language models (LLMs) such as GPT-4o, we were able to generate diverse and controlled examples of biased and unbiased sentences. We generated these sentences as example comments present in the model, data, and dataspace cards of two use cases, highlighting if bias is present in the model or data, which are explained below:

*Customer interaction dataset* In contact centre environments, agents engage with customers across multiple communication channels such as voice calls, chat sessions, and emails. These interactions are often transcribed and analysed to extract insights related to customer sentiment or agent performance. Real-time classification of these transcripts can facilitate immediate feedback mechanisms, such as detecting customer emotions or evaluating agent responses. Importantly, this data also offers an opportunity to assess potential biases exhibited by agents during customer interactions. By detecting biased patterns in interaction data, organisations can implement targeted interventions such as bias-awareness training, thereby promoting equitable treatment across all customer segments.

*Customer data uploaded by organisations* Organisations often integrate customer data from third-party sources such as Customer Relationship Management (CRM) or Enterprise Resource Planning (ERP) systems into their operational pipelines. These datasets may include survey responses or customer profiles provided in an unverified, "as-is" format. Such data is susceptible to hidden biases, especially when the data is selectively sampled or reflects limited demographic representation. When machine-learning models are trained on such biased inputs, they can propagate these underlying biases into model training and ultimately skew the predictions. This presents a significant challenge to ensuring fairness and accuracy in data-driven decision-making processes.

For each use case, 50 biased and 50 non-biased comment sentences were generated and internally reviewed for label verification. These datasets were then used to fine-tune the pre-trained transformer-based models for better textual bias detection. However, synthetic data generation is not without challenges. Overfitting to synthetic patterns, limited diversity, and the potential replication of existing biases are notable concerns. Therefore, while synthetic data can supplement real-world datasets, it should be used judiciously, ensuring that models trained on such data generalise well to authentic scenarios.

**Table 1** Descriptive analysis: neural media bias detection using distant supervision with BABE

Labels	Sub-Group 1 (8 Annotators)	Sub-Group 2 (5 Annotators)
Biased	800	1863
Unbiased	746	1810
No-agreement	154	1
Total sentences	1700	3674

**Table 2** Descriptive analysis: HuggingFace comments—manual annotation

Label	Count	Total
Bias	41	69
Non-bias	28	

### 2.3.3 Hugging face model card comments and manual annotations

To obtain real-world examples of biased language in model documentation, we scraped comments from Hugging Face Model Cards using two sets of keywords:

- **Set 1 (Generic Bias Terms):** *"bias", "biases"*
- **Set 2 (Specific Bias Indicators):** *"overrepresent", "viewpoint", "underrepresent", "stereotypes", "personal information", "hateful", "abusive", "violent", "discriminatory", "prejudicial", "sexual", "errors", "incorrect", "factual", "irrelevant", "repetitive", "consciousness"*

The scraping process yielded 569 links from Set 1 and 217 links from Set 2. However, many of these links were duplicates, as each keyword generated a separate link, and many model cards were children of a common parent. Additionally, a significant number of cards contained headings without accompanying information, limiting the availability of substantive content for analysis. Thus, we got 69 distinct sentences for the HuggingFace model cards.

From the collected data, we manually annotated 69 sentences for bias. Three annotators (Labellers 1, 2, and 3) independently labelled the sentences as "biased" or "non-biased". Labellers 2 and 3 demonstrated high agreement, with a Cohen's Kappa score of 0.94 and only 17 disagreements. Labeller 1's annotations, however, did not align with those of Labellers 2 and 3. To resolve these discrepancies, a fourth annotator (Labeller 4) acted as a tiebreaker for the 17 contentious sentences. The final distribution of annotations is presented in Table 2.

This manually annotated dataset provides valuable insights into the manifestation of bias in model documentation and serves as a critical resource for training and evaluating bias detection models in real-world contexts.

## 3 Results and discussion

This section details the results achieved across all datasets and offers a discussion on the insights drawn from the experimental analysis.

**Table 3** Performance comparison of models on SG1 and SG2

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Macro F1 (%)
XLNet	78.64	78.66	78.26	77.86	78.56
ELECTRA	81.46	81.73	81.25	80.81	81.36
DistilBERT	76.67	78.98	76.43	75.19	76.12
RoBERTa	78.59	79.78	78.78	77.85	78.39
(a) SG1 Results					
XLNet	79.69	79.65	79.50	78.95	79.64
ELECTRA	81.08	81.45	80.97	80.52	81.00
DistilBERT	77.52	77.94	77.37	76.85	77.40
RoBERTa	80.10	80.79	80.10	79.47	79.95
(b) SG2 Results					

### 3.1 Neural media bias detection using distant supervision with BABE

The comparative analysis of transformer-based models on the BABE datasets with sub-groups (SG1 and SG2) reveals consistent trends in model performance across varying levels of annotation agreement. The ELECTRA model emerges as the most effective model, achieving the highest performance (above 80% score across all metrics) in both sub-groups, indicating its strong capacity to detect subtle linguistic patterns of bias, highlighted in Table 3. RoBERTa also delivers competitive results, particularly in SG2. While all models achieve slightly higher scores on SG2, this may be due to differences in sentence content or annotation characteristics. SG2, for instance, contains fewer "No Agreement" cases, which could reflect clearer guidelines or more consistent annotator interpretations. In contrast, SG1, with a larger number of annotators and more instances of disagreement, appears to pose a greater challenge for model generalisation. DistilBERT consistently trails the other models, indicating that smaller architectures may struggle with the subtlety required for bias detection. Overall, the results highlight the interplay between model capacity and the properties of annotated data in achieving reliable textual bias detection.

### 3.2 Synthetic sentence generation using ChatGPT

The results from the synthetic dataset experiments show that all the Transformer models achieved high performance, tabulated in Table 4. When trained and tested on the *Customer Organisation* dataset, ELECTRA achieves the highest performance score, followed by XLNet, while RoBERTa and DistilBERT show slightly weaker performance. In contrast, when models are applied to the *Customer Interaction* dataset, all models exhibit near-perfect scores, with DistilBERT achieving 100% performance score across all metrics. Since the dataset was created using ChatGPT (4o),

**Table 4** Bias detection performance with synthetic data from Customer Organisation and Customer Interaction datasets in percentage (%) fine-tuned on ChatGPT data

Train and test dataset	Model	Accuracy	F1-score	Recall	Precision
Customer Organisation	XLNET	96	96	96	96
	ELECTRA	98	98	98	98
	RoBERTa	91	91	92	92
	Distil-BERT	92	92	92	92
Customer Interaction	XLNET	99	99	99	99
	ELECTRA	98	98	98	98
	RoBERTa	98	98	98	98
	Distil-BERT	100	100	100	100

which itself is based on transformer architectures trained on large corpora, it is plausible that the generated text contains stylistic patterns, lexical regularities, or stereotypical expressions of bias that closely align with the distributions already embedded in the pre-trained transformer models. As a result, the generated sentences may include certain patterns, keywords, or styles that these models already "know" from pre-training. These results highlight the importance of evaluating synthetic data not only in terms of label accuracy but also in terms of its linguistic and contextual richness. While synthetic datasets can be useful for rapid prototyping and controlled testing, they must be carefully assessed for representational fidelity to ensure that models trained on them generalise to real-world bias detection scenarios.

### 3.3 Hugging Face model card comments

The results, tabulated in Table 5, clearly demonstrate the strength of a cascaded fine-tuning approach for bias detection, where transformer models are progressively fine-tuned using multiple datasets in sequence. Starting with a publicly available dataset (BABE), followed by synthetic task-specific samples from ChatGPT (Customer Interaction dataset), and finally 80% of in-domain real-world comments (from Hugging Face), the models gain increasing task alignment and contextual sensitivity at each stage. This setup of moving from broad to narrow domain relevance learning allows the models to build on general knowledge before specialising in real-world bias detection, leading to significantly improved performance in the full fine-tuning setting (using few samples from Hugging Face model cards) across all metrics.

In the partial fine-tuning setting, where models receive no task-specific fine-tuning, performance drops substantially, particularly for lighter models like DistilBERT, highlighting the necessity of domain adaptation in bias detection tasks. For XLNet, ELECTRA, and DistilBERT, the cascade involving all three datasets leads to strong and nearly identical final performance (mid-80 s or higher scores), confirming that gradual fine-tuning across increasingly relevant datasets supports effective transfer learning. The DistilBERT model achieved the most significant improvement in classification accuracy using the proposed cascaded fine-tuning approach, with a 32.09% increase for *Customer*

**Table 5** Performance comparison of XLNet, ELECTRA, DistilBERT, and RoBERTa models under Partial fine-tuned and Fully fine-tuned cascaded training setups for customer interaction and organisational data

Model	Training technique	Accuracy	Precision	Recall	F1-score	AUC PR
Customer interaction dataset						
XLNet	FF	85.75%	88.57%	85.71%	85.07%	85.23%
XLNet	PF	62.32%	61.12%	62.32%	60.85%	72.22%
ELECTRA	<b>FF</b>	<b>85.71%</b>	<b>88.57%</b>	<b>85.71%</b>	<b>85.08%</b>	<b>98.61%</b>
ELECTRA	PF	65.22%	66.82%	65.22%	65.53%	77.50%
DistilBERT	<b>FF</b>	<b>85.71%</b>	<b>88.57%</b>	<b>85.71%</b>	<b>85.08%</b>	<b>98.61%</b>
DistilBERT	PF	53.62%	53.62%	53.62%	53.62%	68.87%
RoBERTa*	FF	92.86%	93.65%	92.86%	95.83%	97.05%
RoBERTa*	PF	62.32%	61.00%	62.32%	59.59%	57.58%
Customer data by organisations dataset						
XLNet	FF	71.43%	72.14%	71.43%	70.16%	83.86%
XLNet	PF	57.97%	55.66%	57.97%	55.32%	69.41%
ELECTRA	<b>FF</b>	<b>85.71%</b>	<b>85.71%</b>	<b>85.71%</b>	<b>85.08%</b>	<b>97.50%</b>
ELECTRA	PF	62.32%	63.31%	62.32%	62.61%	79.65%
DistilBERT	FF	85.71%	88.57%	85.71%	85.71%	97.05%
DistilBERT	PF	62.32%	61.12%	62.32%	60.85%	75.47%
RoBERTa*	FF	85.71%	88.57%	85.71%	85.08%	95.26%
RoBERTa*	PF	52.17%	49.41%	52.17%	49.95%	68.89%

RoBERTa did not perform well on the BABE combined sub-group dataset, affecting model performance in cascade settings. Thus, skipped the BABE-based fine-tuning. Here, PF means Partial Fine-tuned while FF means Fully Fine-tuned

*Interaction* and a 23.39% increase for *Customer Data by Organisation* dataset.

The RoBERTa model performed particularly poorly on the BABE dataset, which affected the overall performance on the Hugging Face comments dataset. Thus, we reported its performance skipping fine-tuning on the BABE dataset. However, it still achieves the strongest performance among all models and highlights the importance of choosing the right sequence and quality of data in cascaded setups, where more is not always better, and strategic dataset ordering can lead to superior outcomes. The performance result of these models on another use case (i.e., Customer data uploaded by organisations) is attached as an appendix table, Table 5.

Overall, these results confirm that cascaded full fine-tuning learning is an effective and adaptable method for bias detection in text, particularly in low-resource settings. They also stress the importance of both dataset relevance and sequencing when designing training pipelines for bias detection tasks.

## 4 Conclusion, limitations and future works

This study evaluates the capacity of four widely used Transformer architectures (XLNet, DistilBERT, RoBERTa, and ELECTRA) to identify bias in AI documentation (Model, Data, and Dataspace Cards). We first benchmarked the models on the publicly available BABE corpus, whose news-article sentences were independently annotated by two experts. All the classification models achieved respectable performance, with ELECTRA surpassing the score of 80% on every metric, thereby establishing a strong in-domain baseline.

As the focus of this study was to detect bias in AI cards, we were unable to find any relevant datasets. Thus, we generated two synthetic datasets with ChatGPT that emulate realistic documentation scenarios for the use cases: *Customer Interaction* and *Customer Data Uploaded by Organisations*. Insights from the synthetic experiments were surprising. When training and testing were performed entirely on the synthetic data, every model achieved above 96% performance score, showing that lexical and stylistic cues supplied by the ChatGPT (a Transformer model) are easily captured. As all the models' performance for the generated dataset was nearly perfect, we scraped the Hugging Face with two sets of keywords to create a real-world comments dataset. Four annotators labelled this dataset in a double-blind setting.

We implemented a cascaded training pipeline where Transformer-based models (pre-trained on the large synthetic set) were fine-tuned on the BABE, then the ChatGPT synthetic dataset, and then fine-tuned and evaluated on the

Hugging Face dataset. Overall, our findings demonstrate that a cascaded full fine-tuning learning training method offers a robust and versatile framework for detecting textual bias, performing especially well when annotated resources are limited.

### 4.1 Limitations

A key limitation of this study lies in the size of the real-world evaluation set collected from Hugging Face. This is largely attributable to how the AI models/data cards are documented. From the total of around 800 original model/data cards, we got only 69 original sentences, due to two reasons: either the models developed were in parent-child configuration, which means that the child inherited the same biases as present in the parent, or the same biases are present in multiple models/data, resulting in repeated identical statements. As a result, the number of unique, bias-relevant sentences available for annotation is modest. However, prior work [30, 31] shows that small, carefully curated bias test sets and low-resource supervised datasets can still reveal stable bias and transfer patterns when combined with robust evaluation and pretrained models. Future studies would benefit from broader community-driven documentation practices that yield more diverse evaluative data.

The relatively small size of the test set further constrains statistical robustness. Minor variations in test samples could lead to observable fluctuations in the reported performance metrics; therefore, the results should be interpreted with appropriate caution.

An additional limitation relates to the sensitivity of cascaded fine-tuning to dataset quality. In preliminary experiments, fine-tuning RoBERTa on the BABE dataset led to degraded classification performance in the cascaded setting. Including these results without contextualisation could incorrectly suggest that RoBERTa is inherently less effective for bias detection. To avoid this misinterpretation, BABE-based fine-tuning was excluded from the final pipeline for RoBERTa. This observation highlights that cascaded training is highly dependent on data relevance and alignment, and that incorporating weakly aligned or domain-mismatched datasets can negatively impact even high-performing models by amplifying error propagation across training stages.

Furthermore, the real-world dataset was constructed using a targeted keyword-based retrieval strategy to capture bias-related content within Hugging Face documentation. This approach ensures relevance to the task and reflects common documentation practices; however, it may not capture all forms of implicitly expressed bias. To address this, synthetically generated data containing implicit bias patterns was incorporated into the cascaded training pipeline,

enabling the models to learn representations that extend beyond explicit lexical cues.

Finally, the relatively small size of the test set constitutes a limitation, as it may introduce variability in the reported performance metrics. Therefore, the results should be interpreted with caution, since minor changes in the test samples could produce noticeable fluctuations in model evaluation outcomes.

## 4.2 Future works

Future research can advance textual bias detection by adopting the following three techniques:

1. Advanced fusion techniques, such as hierarchical late fusion or gated multimodal fusion, can integrate heterogeneous evidence (e.g., lexical cues, syntactic patterns, metadata, and discourse structure). This technique makes the model reason over multiple, complementary signals instead of relying on surface-level word associations. This strategy is particularly effective for detecting implicit bias, as it enables the classifier to aggregate numerous weak cues that, individually, would fall below the decision threshold.
2. Cross-attention mechanisms can be used to align candidate sentences with an external knowledge source, for example, a taxonomy of bias types or a curated list of protected attributes, such as gender, race or ethnicity, nationality, religion, sexual orientation, disability, age, socioeconomic class, or political affiliation. By letting the model attend jointly to the tokens of the input sentence and the bias descriptors (the tokens in the protected attribute list), the cross-attention model can (a) detect implicit references that would be missed by keyword matching alone and (b) ground its predictions in an interpretable set of bias triggers.
3. A dual-encoder architecture with an explicit bias definition branch can disentangle the semantics of the candidate sentence from the semantics of the bias taxonomy. One encoder specialises in representing the sentence; the other encodes formal bias definitions or exemplar phrases. Their joint embedding space then supports fine-grained similarity scoring, allowing the system to pinpoint not only whether bias is present but also which specific subtype (e.g., stereotyping, exclusion, unequal framing) is triggered.

Together, these three directions promise to move bias detection from shallow pattern matching toward deeper, knowledge-aware reasoning, thereby improving robustness, transferability, and explanatory qualities, which are

essential for safe deployment in real-world AI documentation pipelines.

**Author Contributions** A.D., A.G., A.O., E.C., E.M., I.U., J.P.M., M.K., S.S., and T.I. contributed equally to this work. All authors jointly contributed to the conceptualisation and design of the study, development of the methodology, analysis and interpretation of results, and writing of the manuscript. All authors reviewed, revised, and approved the final manuscript. The author's names appear in alphabetical order.

**Funding** Open Access funding provided by the IReL Consortium. This publication has emanated from research conducted with the financial support of Research Ireland under Grant Numbers 12/RC/2289\ P2 - Insight Research Ireland Centre for Data Analytics and under Grant Number 20/SP/8955 - School of Computer Science. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The full funding declaration will be added in the final version.

**Data Availability** The datasets used in this study can be provided for research purposes only upon request.

## Declarations

**Ethical approval** This study used only publicly available datasets (BABE corpus, Hugging Face documentation comments) and synthetically generated text, with no personal or sensitive information involved. Human annotation was carried out by independent labellers with informed consent, and no identifying data were collected. The research adheres to GDPR requirements and the EU Ethics Guidelines for Trustworthy AI.

**Conflict of interest** The authors declare no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Raza, S., Garg, M., Reji, D.J., Bashir, S.R., Ding, C.: Nbias: a natural language processing framework for bias identification in text. *Expert Syst. Appl.* **237**, 121542 (2024)
2. Cui, H., Yasseri, T.: Ai-enhanced collective intelligence. *Patterns* **5**(11), 101074 (2024)
3. Gardazi, N.M., et al.: Bert applications in natural language processing: a review. *Artif. Intell. Rev.* **58**, 1–49 (2025)
4. Donald, A., et al.: A semantic approach for linked model, data, and dataspace cards. *IEEE Access* **13**, 110194–110207 (2025)

5. Mitchell, M., et al.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229 (2019)
6. Bevara, R.V.K., Mannuru, N.R., Karedla, S.P., Xiao, T.: Scaling implicit bias analysis across transformer-based language models through embedding association test and prompt engineering. *Appl. Sci.* **14**, 3483 (2024)
7. Raza, S., Reji, D.J., Ding, C.: Dbias: detecting biases and ensuring fairness in news articles. *Int. J. Data Sci. Anal.* **17**, 39–59 (2024)
8. Dantas, P.V., da Silva Jr, W.S., Cordeiro, L.C., Carvalho, C.B.: A comprehensive review of model compression techniques in machine learning. *Appl. Intell.* **54**, 11804–11844 (2024)
9. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey. *Comput. Linguist.* **50**, 1097–1179 (2024)
10. Cortiz, D.: Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. arXiv preprint [arXiv:2104.02041](https://arxiv.org/abs/2104.02041) (2021)
11. Clark, K., Luong, M.-T., Le, Q.V. Manning, C.D.: Electra: pre-training text encoders as discriminators rather than generators. arXiv preprint [arXiv:2003.10555](https://arxiv.org/abs/2003.10555) (2020)
12. Yang, Y., Duan, H., Abbasi, A., Lalor, J.P., Tam, K.Y.: Bias a-head? Analyzing bias in transformer-based language model attention heads, pp. 276–290 (2025)
13. Goyal, M., Mahmoud, Q.H.: A systematic review of synthetic data generation techniques using generative AI. *Electronics* **13**, 3509 (2024)
14. González-Sendino, R., Serrano, E., Bajo, J.: Mitigating bias in artificial intelligence: fair data generation via causal models for transparent and explainable decision-making. *Futur. Gener. Comput. Syst.* **155**, 384–401 (2024)
15. Long, L., et al.: On llms-driven synthetic data generation, curation, and evaluation: a survey (2024). arXiv preprint [arXiv:2406.15126](https://arxiv.org/abs/2406.15126)
16. Spliethöver, M., et al.: Adaptive prompting: Ad-hoc prompt composition for social bias detection (2025). arXiv preprint [arXiv:2502.06487](https://arxiv.org/abs/2502.06487)
17. Wei, X., Kumar, N., Zhang, H.: Addressing bias in generative ai: challenges and research opportunities in information management. *Inf. Manag.* **62**(2), 104103 (2025)
18. Chen, B., Zhang, Z., Langrené, N., Zhu, S.: Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Patterns* (2023)
19. Lin, Z., et al.: Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artif. Intell. Rev.* **57**, 243 (2024)
20. Li, M., et al.: Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks (2025). arXiv preprint [arXiv:2502.04419](https://arxiv.org/abs/2502.04419)
21. Yang, Z., et al.: Xlnet: generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **32** (2019)
22. Areshey, A., Mathkour, H.: Exploring transformer models for sentiment classification: a comparison of bert, roberta, albert, distilbert, and xlnet. *Expert. Syst.* **41**, e13701 (2024)
23. Chauhan, A., Mohana, R.: Combining transfer and ensemble learning models for image and text aspect-based sentiment analysis. *Int. J. Syst. Assur. Eng. Manag.* **16**(3), 1–19 (2025)
24. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach (2019). arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692)
25. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019). arXiv preprint [arXiv:1910.01108](https://arxiv.org/abs/1910.01108)
26. Multimodal Tokenizer. [Available Online] [https://huggingface.co/docs/transformers/main\\_classes/tokenizer](https://huggingface.co/docs/transformers/main_classes/tokenizer). Accessed 22 May 2025
27. Transformers. [Available Online] <https://huggingface.co/docs/transformers/en/index>. Accessed 22 May 2025
28. Babe - media bias dataset: Annotations by experts. [Available Online] <https://www.kaggle.com/datasets/timospinde/babe-media-bias-annotations-by-experts>. Accessed 29 May 2024
29. Spinde, T., et al.: Neural media bias detection using distant supervision with babe–bias annotations by experts (2022). arXiv preprint [arXiv:2209.14557](https://arxiv.org/abs/2209.14557)
30. Nagar, K.L., et al.: Bias detection using textual representation of multimedia contents 408–416 (2023)
31. Steinert, S., et al.: A refined approach for evaluating small datasets via binary classification using machine learning. *PLoS ONE* **19**, e0301276 (2024)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.