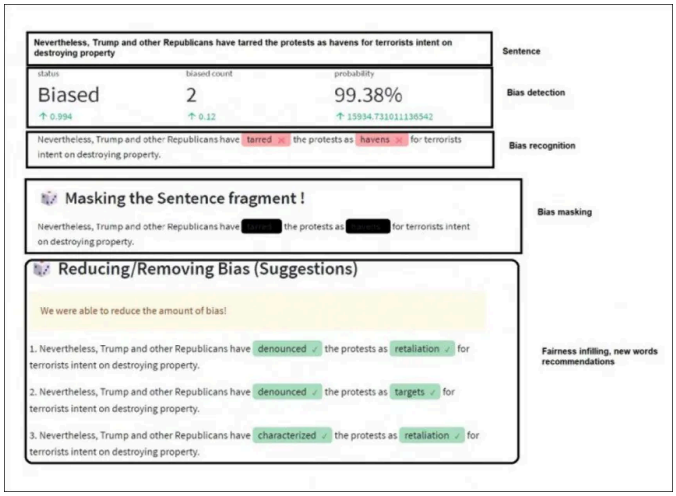
An AI-Driven Bias Checker for News Articles, Available in Python

Originally published at https://www.unite.ai/an-ai-driven-bias-checker-for-news-articles-available-in-python/

Researchers in Canada, India, China and Australia have collaborated to produce a freely-available Python package that can effectively be used to spot and replace 'unfair language' in news copy.

The system, titled *Dbias*, uses various machine learning technologies and databases to develop a three-stage circular workflow that can refine biased text until it returns a non-biased, or at least more neutral version.

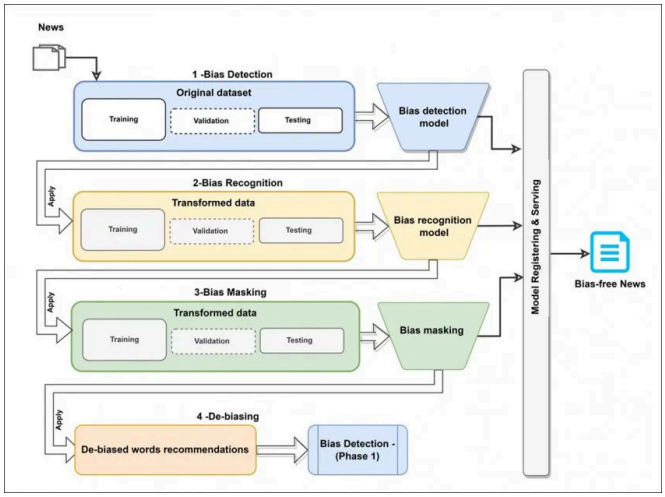


Loaded language in a news snippet identified as 'biased' is transformed into a less incendiary version by Dbias. Source: https://arxiv.org/ftp/arxiv/papers/2207/2207.03938.pdf

The system represents a reusable and self-contained pipeline that can be installed via Pip from Hugging Face, and integrated into existing projects as a supplemental stage, add-on, or plugin.

In April, similar functionality implemented in Google Docs came under criticism, not least for its lack of editability. Dbias, on the other hand, can be more selectively trained upon any corpus of news that the end-user wishes, retaining the ability to develop bespoke fairness guidelines.

The critical difference is that the Dbias pipeline is intended to automatically transform 'loaded language' (words that add a critical layer to factual communication) into neutral or prosaic language, rather than to school the user on an ongoing basis. Essentially, the end-user will define ethical filters and train the system accordingly; in the Google Docs approach, the system is – arguably– training the user, in a unilateral fashion.



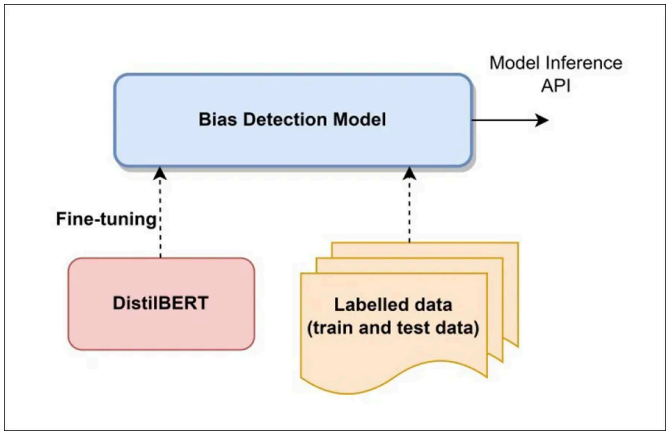
Conceptual architecture for the Dbias workflow.

According to the researchers, Dbias is the first truly configurable bias detection package, in contrast to the off-the-shelf assembly projects that have characterized this sub-sector of Natural Language Processing (NLP) to date.

The new paper is titled *An Approach to Ensure Fairness in News Articles*, and comes from contributors at the University of Toronto, Toronto Metropolitan University, Environmental Resources Management at Bangalore, DeepBlue Academy of Sciences in China, and The University of Sydney.

Method

The first module in Dbias is *Bias Detection*, which leverages the DistilBERT package – a highly optimized version of Google's fairly machine-intensive BERT. For the project, DistilBERT was fine-tuned on the Media Bias Annotation (MBIC) dataset.



MBIC consists of news articles from a variety of media sources, including the Huffington Post, USA Today and MSNBC. The researchers used the extended version of the dataset.

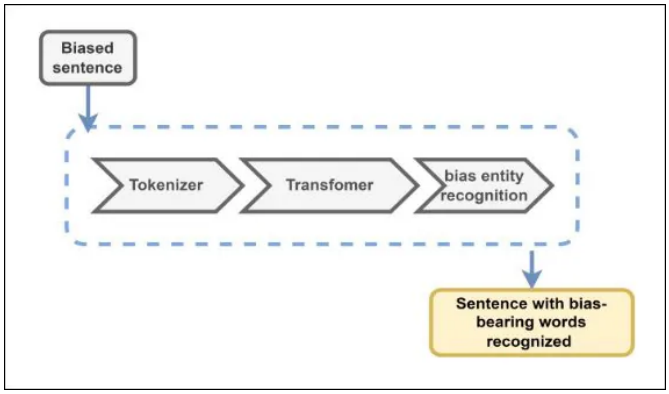
Though the original data was annotated by crowdsourced workers (a method which came under fire late in 2021), the researchers of the new paper were able to identify additional unlabeled instances of bias in the dataset, and appended these manually. The identified incidences of bias related to race, education, ethnicity, language, religion, and gender.

The next module, *Bias Recognition*, uses Named Entity Recognition (NER) to individuate biased words from the input text. The paper states:

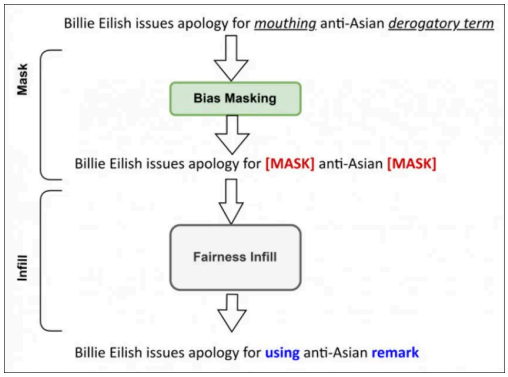
> *'For example, the news "Don't buy the pseudo-scientific hype about tornadoes and climate change" has been classified as biased by the preceding bias detection module, and the biased recognition module can now identify the term "pseudo-scientific hype" as a biased word.'*

NER is not specifically designed for this task, but has been used before for bias identification, notably for a 2021 project from Durham University in the UK.

For this stage, the researchers used RoBERTa combined with the SpaCy English Transformer NER pipeline.



The next stage, *Bias Masking*, involves a novel multiple-mask of the identified bias words, which operates sequentially in cases of multiple identified bias words.



Loaded language is replaced with pragmatic language in the third stage of Dbias. Note that 'mouthing' and 'using' equate to the same action, though the former is considered derisive.

As necessary, the feedback from this stage will be sent back to the beginning of the pipeline for further evaluation until a number of suitable alternative phrasings or words have been generated. This stage uses Masked Language Modeling (MLM) along lines established by a 2021 collaboration led by Facebook Research.

Normally the MLM task will mask 15% of the words randomly, but the Dbias workflow instead tells the process to take the identified biased words as input.

The architecture was implemented and trained on Google Colab Pro on a NVIDIA P100 with 24GB of VRAM at a batch size of 16, using just two labels (*biased* and *unbiased*).

Tests

The researchers tested Dbias against five comparable approaches: LG-TFIDF with Logistic Regression and TfidfVectorizer (TFIDF) word embeddings; LG-ELMO; MLP-ELMO (a feed-forward artificial neural network containing ELMO embeddings); BERT; and RoBERTa.

Metrics used for the tests were accuracy (ACC), precision (PREC), recall (Rec) and an F1 score. As the researchers had no knowledge of any existing system that could accomplish all three tasks in a single pipeline, dispensation was made for the competing frameworks, by evaluating only Dbias's primary tasks – bias detection and recognition.

Model	PREC	REC	F1
LG-TFIDF	62%	61%	62%
LG- ELMO	67%	69%	68%
MLP- ELMO	69%	68%	68%
Bert	72%	69%	71%
RoBERTa	76%	70%	73%
Our approach	**77%**	**74%**	**75%**

Results from the Dbias trials.

Dbias succeeded in surpassing results from all competing frameworks, including those with a heavier processing footprint

The paper states:

'The result also shows that deep neural embeddings, in general, can outperform traditional embedding methods (e.g., TFIDF) in the bias classification task. This is shown by the better performance of deep neural network embeddings (i.e., ELMO) compared to TFIDF vectorization when used with LG.

'This is probably because deep neural embeddings can better capture the context of the words in the text in different contexts. The deep neural embeddings and deep neural methods (MLP, BERT, RoBERTa) also perform better than traditional ML method (LG).'

The researchers also note that Transformer-based methods outperform competing methods in bias detection.

An additional test involved a comparison between Dbias and various flavors of the SpaCy Core Web, including core-sm (small), core-md (medium), and core-lg (large). Dbias was able to lead the board also in these trials:

Model	PREC	REC	F1	ACC
Core-sm	59%	27%	37%	37%
Core-md	61%	45%	52%	53%
Core-lg	60%	62%	60%	67%
Our approach	**66%**	**65%**	**63%**	**72%**

The researchers conclude by observing that bias recognition tasks generally show better accuracy in larger and more expensive models, due – they speculate – to the increased number of parameters and data points. They also observe that the efficacy of future work in this field will depend on greater efforts to annotate high-quality datasets.

The Forest and the Trees

Hopefully this kind of fine-grained bias recognition project will eventually be incorporated into bias-seeking frameworks that are able to take a less myopic view, and to take into consideration that choosing to cover any particular story is in itself an act of bias that's potentially driven by more than just reported viewing statistics.

First published 14th July 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More