VECTOR INSTITUTE

Search Search...

Insights

Insights

- Vector Insights
- Research Insights
- AI Trust and Safety Principles
- Newsroom
- AI Ecosystem Reports
- IP For AI Innovations

- About

About

About Vector

- Overview
- Team and Leadership
- Careers at Vector
- Annual Reports
- Alumni

Sign In

Sign In

This is a navigation window which overlays the main content of the page. Pressing the "X" at the top right corner of the modal will close the modal and bring you back to where you were on the page.

See posts about

- Announcement
- Blog
- Case Study
- Generative AI
- Health
- Insights
- Intellectual Property
- Masters Programs
- News
- Partnership
- Press

- Program
- Research
- Scholarship
- Success Stories
- Trustworthy AI

This is a call to action window which overlays the main content of the page. Pressing the "X" at the top right corner of the modal will close the modal and bring you back to where you were on the page.

Let's Collaborate

Partnership Sector

First Name

Last Name

Company

Company Email

City

Country

Please select the country you are located in

Tell us about your organization

☐ Yes, contact me

By submitting this form you consent to the given information be used by Vector to contact you about my inquiry. You have read and agreed to the Privacy Policy

Send Message

Neutralizing Bias in AI: Vector Institute's UnBIAS Framework Revolutionizes Ethical Text Analysis

December 5, 2023

Insights Research Trustworthy AI

By Mark Coatsworth and Shaina Raza

In today's information age, accurate and bias-free content are paramount. AI systems have already become instrumental in disseminating information, increasing the potential for biased and incorrect information. Biased training data can lead to algorithms that perpetuate stereotypes and reinforce biases, and biased algorithms can lead to the spread of false information. This has broad implications

across the news media, social networks, regulatory compliance, governance, and other policy matters, increasing the potential for significant harm.

To tackle this bias and promote the ethical use of large language models (LLMs), the Vector Institute's AI Engineering team built **UnBIAS** (unbias-mkdocs.readthedocs.io), a cutting-edge text analysis framework that assesses and corrects biases in textual data through bias classification, named entity recognition for bias identification, and text debiasing.

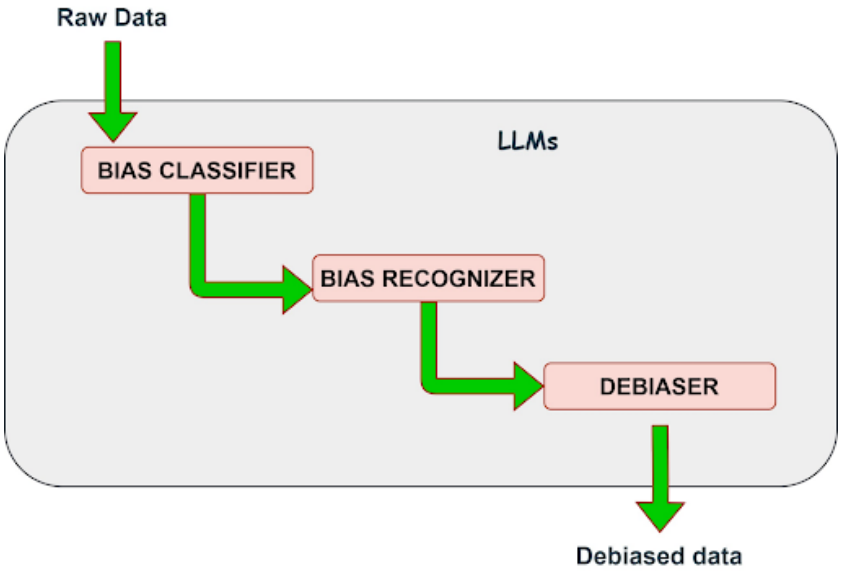
With so much misinformation, disinformation, and bias common in the areas of news, public policy, and regulations, digital media cannot possibly be managed via conventional means. The creation of UnBIAS, led by Shaina Raza, an Applied ML Scientist in Responsible AI at Vector, provides an AI-backed framework to identify bias in communications and a simple means to replace biased text with neutral, unbiased content.

How does it work?

Based on a Python library developed by Vector and released as an open-source library, UnBIAS leverages LLMs as its foundation for detecting biased content. Users provide the model with text (either single or a batch of textual content), for example *"Men are naturally better than women at sports."* The integrated classifier model examines that text to determine if bias is present. If detected through binary classification, the model returns a confidence score and replacement text, such as *"Men are naturally better than women at sports. BIASED(95%)"* Next, the confidence score and the sequence advances to the token classifier stage. This token classifier is adept at identifying and flagging tokens of biases, both blatant and subtle, extending up to n-grams tokens within the text.

The final stage involves the **debiaser**: a utility to replace the previously biased text with new, unbiased text. The debiaser would replace our previous example with the text: *"Individuals of different genders can excel in sports based on their unique skills, training, and dedication rather than gender alone."*

The debiaser employs instruction-based fine-tuning, augmented with both parameter-efficient strategies and 4-bit quantization techniques (allowing the model to run even without heavy computation) to neutralize the biased texts. When debiasing a statement, the goal is not to change the original meaning, but rather to present it in a way that is free from bias and stereotypes. Each mode in the resulting UnBIAS toolkit is fine-tuned for efficient inference and arranged sequentially to form a streamlined pipeline architecture.



Dataset Preparation

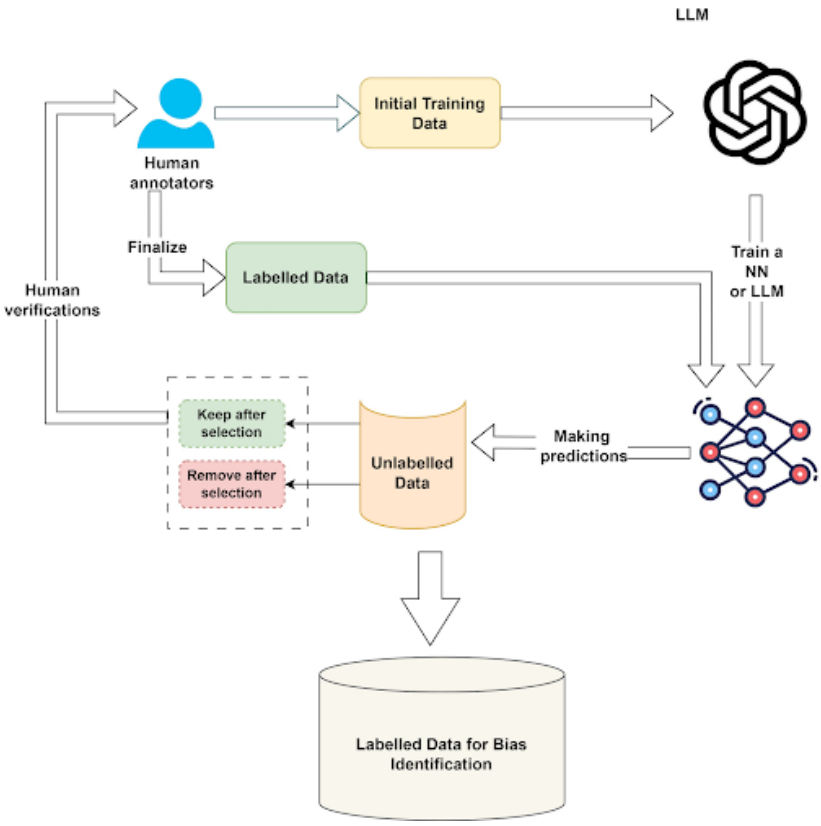
In developing the Python library, Raza and the AI Engineering team ran up against a major hurdle: a lack of high-quality, open source training data that could be labeled for bias identification. To solve for this, they curated unique datasets of news articles marked for identifying biases; they also created a debiased version of the biased data for fine-tuning the debiaser, as LLMs trained on these datasets are adept at bias detection and rectification.

The datasets were released under open-source licenses:

- Fake News Elections 2024
- News Bias Full Data

The goal is to seamlessly integrate the UnBIAS framework with recommendation systems and other information retrieval software (hiring apps, documents stores, news websites) to produce unbiased results. To this end, the team emphasizes the importance of high-quality, diverse data representing a number of bias aspects (gender, age, religion, sexism) to ensure that the dataset works for a wide range of bias identification tasks. Consistency in labeling biased subjective matter was also a challenge, while privacy and ethical considerations necessitated safeguarding individuals' and organizations' confidentiality mentioned in news.

Ensuring that datasets and models are findable, accessible, interoperable, and reusable — the FAIR principles — is a cornerstone of our approach to data and model preparation. Datasets are prepared using AI-based labeling and verified through multiple human-in-the-loop iterations where humans rank, identify, and evaluate biased output, providing a far stronger detection basis than could achieved from regular model training.



What is the Societal Impact?

This toolkit champions ethical AI, aligning with ideals of inclusivity and equity. It is designed to counteract biases across various spectrums, including politics, race, gender, and age, while also addressing misinformation in diverse areas, like climate change and global politics. It is aligned with ethical AI ideals, including inclusivity and equity. It also reflects a broader commitment to continuously aligning Vector's work with AI community values while vigilantly assessing any unintended AI repercussions.

What's Next?

The UnBIAS framework represents a significant advancement in text analysis and bias correction. By utilizing a combination of sophisticated classifiers, and innovative debiasing techniques, UnBIAS addresses the urgent need for accurate and unbiased information dissemination in today's digital age. The framework not only detects and addresses biases in textual content but also preserves the original intent and meaning, thereby promoting fairness and ethical AI practices. This approach aligns with the principles of inclusivity and equity, and holds strong potential to significantly mitigate the spread of bias and misinformation.

Raza and the AI Engineering team hope to make UnBIAS accessible and user-friendly for both data science professionals and non-data science individuals to easily detect, and if necessary mitigate, biases in their documents and texts. They are currently at work expanding the toolkit with more extensive training and evaluation tools, enhancing the framework's capabilities to identify biases across a wide range of domains, including legal, medical, e-commerce, and more.

Related:

Generative AI

Insights

Research

Introducing FlexModel: Breakthrough Framework for Unveiling the Secrets of Large Generative AI Models

Insights

Research

Vector researchers presenting more than 65 papers at NeurIPS 2023

Health

Insights

Safe AI implementation in health: Why the right approach matters