International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 5, Issue 4, 88-98, 2024
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V5I4P109

Original Article

Bias Detection in AI Models Using Word Embedding Association Tests (WEAT)

Khaja Kamaluddin
Masters in Sciences, Fairleigh Dickinson University, Teaneck, NJ, USA, Aonsoft International Inc, 1600 Golf Rd, Suite 1270, Rolling Meadows, Illinois, 60008 USA.

Abstract *- Bias in artificial intelligence (AI) systems has emerged as a critical concern, particularly in natural language processing (NLP), where pretrained word embeddings often encode and amplify societal stereotypes. The Word Embedding Association Test (WEAT) has become a foundational method for detecting such biases by quantifying the associative relationships between conceptually relevant word groups. This review provides a comprehensive synthesis of research and applications related to WEAT and its methodological extensions, covering developments. It begins with a technical overview of static and contextual word embeddings and outlines how various forms of bias manifest within them. We explore the origins and mathematical framework of WEAT, followed by its adaptations such as SEAT and ROME that address limitations in modern transformer-based models. The article then examines practical applications of WEAT in sentiment analysis, recommendation systems, healthcare diagnostics, and legal risk assessments, highlighting its role in AI governance and ethical auditing. A comparative analysis of prominent WEAT-supporting tools and empirical case studies further illustrate its effectiveness and limitations. Finally, we discuss unresolved challenges related to model contextuality, definitional ambiguity, and cross-disciplinary integration. Through this review, we underscore the importance of embedding-level bias audits and advocate for the evolution of WEAT into a more context-aware and policy-aligned framework for responsible AI.*

Keywords *- Word Embedding Association Test (WEAT), bias detection, natural language processing (NLP), contextual word embeddings, algorithmic fairness, ethical AI governance.*

1. Introduction

Artificial intelligence (AI) has become deeply integrated into decision-making systems across various sectors, including healthcare, finance, criminal justice, and education [1]. These systems are often perceived as neutral and objective, yet numerous studies and high-profile incidents have revealed that AI can reflect, reinforce, and even exacerbate existing societal biases [2]. From predictive policing systems disproportionately targeting minority communities to resume-screening algorithms favoring male applicants over equally qualified female candidates, the consequences of biased AI can be both far-reaching and harmful. These issues have catalyzed a growing movement within the research community focused on algorithmic fairness and transparency, aiming to identify, measure, and mitigate the presence of bias in AI systems.

Word embeddings dense-vector representations of language are just one of the numerous elements of the modern AI to which natural language processing (NLP) applications find a foundation [3]. Word2Vec, GloVe and fastText embeddings are algorithms trained on substantial texts to reflect the semantic similarity of words. Nonetheless, such representations usually reflect and capture the statistical biases present in their underlying data. As an example, embeddings can encode such analogies as women are to homemakers as computer programmers are to man, an expression of gender stereotype in web texts or historical literature [4]. By spreading to downstream NLP applications, such fine-grained but widespread associations may adversely affect downstream applications, including sentiment analysis, machine translation, or chatbots in a manner that is unintended and discriminatory.

To make the systematic identification of such biases in the set of word embeddings, researchers have come up with several diagnostic tools. The Word Embedding Association Test (WEAT) is one of the most powerful techniques that based on psychological approaches such as Implicit Association Test (IAT) are used to identify relationships between the target concepts in terms of male versus female names paired with a set of attributes such as career versus family-related words[5]. WEAT measures the degree of bias within vector spaces that can then be described using statistical measures, and as such offers a formal and auditable method of auditing pre-trained language models. The ethical imperative for fairness in AI, as emphasized [6] , highlights the increasing society requirement to have clear algorithm responsibility. What is more, the reason why Cloud Security Posture Management (CSPM) calls out the necessity of automated policy enforcement and audit in safe cloud [7] values relates well to the objectives of bias detection in AI models. Collectively, these views establish a strong rationale behind a systematic review of using WEAT as a critical instrument in the creation of responsible AI.

Figure 1 below is a conceptual representation of the way that biases can creep in and circulate using a typical NLP pipeline data ingestion, and embedding generation on to model deployment as well as where the key intervention or auditing could take place.

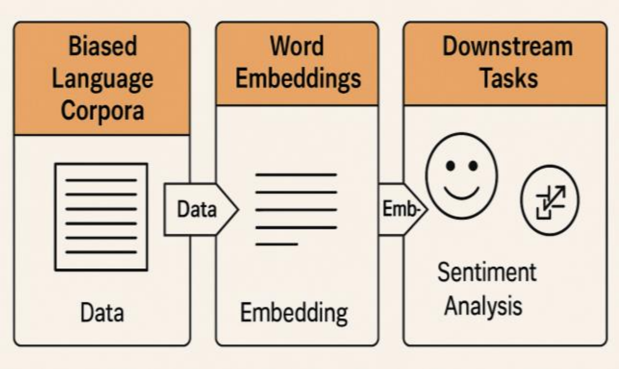

Fig 1: Conceptual Visualization of Bias Propagation in NLP Pipelines

As the reliance on pre-trained embeddings and large language models continues to grow, the importance of such diagnostic techniques cannot be overstated. This review paper is devoted to the biases detection in the AI models through WEAT and methodological extensions thereof, paying attention to developments in the research filed. It gives an elaborate analysis of the operation of WEAT, the comparison with similar methods like SEAT (Sentence Encoder Association Test) are many, and the practical applications of WEAT in the real society in application areas such as healthcare, governance, and control of content are many. Additionally, the review also discusses the tools that apply WEAT, like IBM AI Fairness 360 and WEFE framework and also reveals their limitations and the ethical questions that are not resolved yet.

We structure this article into eight sections. Section 2 outlines the technical background of word embeddings and the types of biases they may encode. Section 3 introduces WEAT and details its core methodology, followed by Section 4, which explores methodological variants and extensions. Section 5 delves into practical applications and domain-specific impacts, while Section 6 surveys the available toolkits for applying WEAT in research and production settings. Section 7 presents empirical results and case studies from the literature, and Section 8 concludes with a discussion on open challenges and ethical implications.

2. Technical Background on Word Embeddings and Bias

Understanding how bias emerges in AI systems requires a foundational grasp of word embeddings the core linguistic structures used by modern NLP models [8]. They are the semantically oriented building blocks of Machine Learning language, the relationships between words regarding their distributional properties. Although they are competent in terms of meaning preservation, they are also, quite likely, to encode and exaggerate biases existing within the training data reflected in the society [9]. In this section, word embeddings and the mechanisms by which they promote bias are explored along with the examples of their unintended use in the real world.

2.1. Word Embedding Fundamentals

Word embeddings are vectorized form of words that enable machine learning models to work with language in a mathematically relevant manner. They form the focus of most NLP applications, including search engines, chatbots, translation services, and sentiment analysis applications. differently to sparse one-hot encoding representations, word embeddings are dense, reflecting semantic connections among words in continuous, lower dimensions.

Three widely adopted embedding models, Word2Vec, GloVe, and fastText share the common goal of learning these vector representations from a large corpora of text [10]. However, they differ in architecture and training mechanisms:

- Word2Vec (Mikolov et al., 2013) uses a shallow neural network to learn word vectors by predicting neighboring words (Skip-gram) or predicting a target word from its context (CBOW).
- GloVe (Global Vectors for Word Representation) relies on matrix factorization of the word co-occurrence matrix, capturing both local and global context.
- Fast Text extends Word2Vec by incorporating subword information, allowing it to create embeddings for rare or out-of-vocabulary words using character n-grams.

All three models follow the distributional hypothesis: words appearing in similar contexts tend to have similar meanings. This principle enables embeddings to capture analogical relationships (e.g., king – man + woman ≈ queen), but it also risks encoding societal biases present in the data.

Table 1: Comparison of Word Embedding Models and Their Vulnerability to Bias

Model	Training Mechanism	Context Sensitivity	Handles OOV Words	Vulnerability to Bias
Word2Vec	Predictive (Skip-gram/CBOW)	Moderate	No	High
GloVe	Matrix factorization	Low	No	High
fastText	Predictive with subwords	Moderate–High	Yes	Moderate

2.2. Definition of Bias in Embeddings

Bias in word embeddings manifests when these vector representations reflect stereotypes or prejudiced associations [11]. While such biases may not be explicitly programmed into a system, they are often inherited from the large-scale textual data used to train the embeddings. Broadly, bias in embeddings can be classified into three types:

- Associative Bias: Occurs when word vectors capture and reinforce stereotypical associations (e.g., "man" is closer to "programmer" than "woman").
- Contextual Bias: It appears in the contextual embedding (Such as BERT) where meaning changes depending on how it is contextualized by other words [12]. This can result in uneven or discrimination depending on the use.
- Representation Bias: Caused by either the underrepresentation or overrepresentation of different groups or concepts to the training data, this introduces bias to the embedding dimension.

Bias such biases may lead to downstream effects associated with high stakes applications. As an example, small changes in the distances between word vectors can distort any similarity measure in sentiment or profiling system or the boundary between class [13]. The dangers associated with all of these biases are not only concerned with ethical barriers, but model layer structural biases can create system-wide inequalities when it comes to predictive decision-makingm [14]. On the same note, the existence of architectural vulnerabilities in cloud-native applications can exacerbate such biases in the event that they are not addressed through mitigation layers when integrating embeddings [15].

2.3. Case Examples of Embedding Bias

Numerous empirical studies have demonstrated that popular word embeddings exhibit troubling patterns of bias. One of the most cited examples is gender bias:

- In classic embeddings trained on web data, analogies such as: "man": "computer programmer": "woman": "homemaker” emerge frequently, reflecting entrenched gender stereotypes.
- Racial or ethnic bias also appears in embeddings, for example, associating certain names or terms with negative sentiment or criminality, particularly in U.S.-based datasets.

These associations can significantly influence applications like resume screening, predictive policing, content moderation, and even medical diagnostics, where neutral treatment is essential. The severity of such bias calls for robust diagnostic methods, of which WEAT is among the most established.

3. Word Embedding Association Test (WEAT): Core Methodology

Improper biases in word embeddings saw growing interest as the need to measure them without relying on ad hoc means was risen [16]. The Word Embedding Association Test (WEAT) has turned out to be one of the most high profile methods to examine prejudice in embedding space. WEAT translates psychological methods applied in the investigation of human cognition namely, the Implicit Association Test (IAT) to a vectorized environment of machine learning. In this section, the set of theoretical backgrounds of WEAT, its mathematical representation and its operational advantages and drawbacks during application of differing models of NLP were presented.

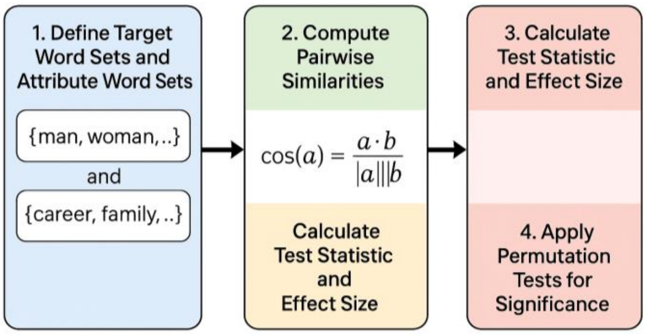
1. Define Target Word Sets and Attribute Word Sets
{man, woman,..}
and
{career, family,..}
2. Compute Pairwise Similarities
$\cos(a) = \frac{a \cdot b}{|a||b}$
Calculate Test Statistic and Effect Size
3. Calculate Test Statistic and Effect Size
4. Apply Permutation Tests for Significance

Fig 2: WEAT Workflow Diagram

3.1. Origins and Theoretical Underpinnings

To make the systematic identification of such biases in the set of word embeddings, researchers have come up with several diagnostic tools [17]. The Word Embedding Association Test (WEAT) is one of the most powerful techniques that based on psychological approaches such as Implicit Association Test (IAT) are used to identify relationships between the target concepts in terms of male versus female names paired with a set of attributes such as career versus family-related words. WEAT measures the degree of bias within vector spaces that can then be described using statistical measures, and as such offers a formal and auditable method of auditing pre-trained language models.

3.2. Mathematical Formulation

The WEAT test requires four sets of words:

- Two target sets (e.g., male names vs. female names)
- Two attribute sets (e.g., career-related words vs. family-related words)

The similarity between each target word and attribute set is computed using cosine similarity. WEAT then calculates a test statistic (S) based on the difference in mean similarity of target words to the two attribute sets.

Formally, for target sets XXX and YYY, and attribute sets AAA and BBB, the WEAT score is defined as:

$$S(X, Y, A, B) = \Sigma\_\{x \in X\}\, s(x, A, B) - \Sigma\_\{y \in Y\}\, s(y, A, B)$$

Where:

$$s(w, A, B) = mean\_\{a \in A\}\, cos(w, a) - mean\_\{b \in B\}\, cos(w, b)$$

In order to check the importance of the perceived association, WEAT employs the permutation test. Multiple random rearrangements are applied between the two sets of target words (typically 10,000; rearrangements) and the resultant test statistics distribution is applied to calculate a p-value. Further, the size of the bias is approximated by computing an effect size. It is, actually, a standardized difference in means (Cohen d), which you may compare on the basis of various experiments and data sets. Such statistical methods, being simple to assemble, provide a potent diagnostic indication in assessing social or semantic biases in vectorized models. The quantum-resistant cryptography as situation, strenuous testing models like, permutation-based evaluations are essential in making sure the measurement is reliable [18] concepts which can be replicated into the issue of AI biasness too.

3.3. Strengths and Limitations of WEAT

WEAT has gained traction as a leading technique due to several compelling advantages:

- Language and Model Agnosticism: It can be applied to any word embedding model that uses vector spaces, regardless of architecture (Word2Vec, GloVe, fastText, etc.)[19].
- Simplicity and Transparency: The methodology of WEAT is quite simple to use, to audit and reproduce, appealing both to the researcher and the practitioner.
- Quantitative Output: WEAT provides a quantitative, as well as a qualitative, impression of bias, not only by supplying the statistical and effect size values but also by performing the analyses in numerical terms.

3.4. However, WEAT also comes with notable limitations:

- Inability to Capture Contextual Bias: WEAT is developed as a static word embedding task, and so it fails to diagnose bias effectively on models such as BERT or GPT, in which the meaning conveyed by a word is dynamically conditioned on context [20].
- Sensitivity to Word List Selection: The results can vary depending on the chosen target and attribute sets, making the test somewhat subjective.

Table 2: Comparison of WEAT, SEAT, and ROME Tests

Method	Target Model Type	Context-Aware	Testing Mechanism	Primary Use Case
WEAT	Static embeddings	No	Cosine similarity + permutation tests	Bias in traditional embeddings
SEAT	Contextual models (BERT, GPT)	Yes	Sentence-level similarity scoring	Bias in transformer encoders
ROME	Transformer LLMs (GPT-2/3)	Yes	Direct model memory editing	Causal analysis and intervention

4. Data Sparsity and Low-Frequency Words

Rare words may yield unstable similarity measures, particularly in smaller or domain-specific corpora. These shortcomings explain the significance of model-conscious assessments and soundly crafted experimental design. The effectiveness of the process of caching training data focuses on the essential importance of factors of data conditioning and representativeness of the samples directly affecting the statistic level of the WEAT significance [21].

4.1. Extensions and Variants of WEAT

While the Word Embedding Association Test (WEAT) gives a solid foundation of how the bias can be measured in the static word embeddings, the further development of the NLP models specifically models based on transformers such as BERT and GPT has brought new issues to the matter. The contextual models provide dynamic embeddings and thus WEAT alone is not enough. To overcome this, a variety of extensions and alternatives to WEAT with modifications of the fundamental principles in order to support language representations that are sensitive to context in order to conduct more finer-grained discrimination are being created. The following are the main developments discussed in this section, such as; SEAT, ROME, and causal probing methods.

4.2. SEAT (Sentence Encoder Association Test)

To apply the idea of WEAT to context-sensitive embeddings, the Sentence Encoder Association Test (SEAT) was added that measures the degree to which a model is context-dependent. Such models do not denote a fixed single vector to each word anymore; rather what they bring is the representation of a word to vary according to text around it. This entails the need of having a testing structure that runs at the sentence level and not at the word level [22]. SEAT has a similar format as WEAT except that it involves the use of full sentences as opposed to individual words. It determines whether there exist sentences incorporating target concepts in neutral wordings that can be found to be more semantically related to specific set of properties than they are to other sets of properties. The CLS token output or pooled representations of the transformer models are usually used to extract sentence embeddings.

For example, to test gender bias, SEAT might compare the embeddings of:

- “She is a brilliant engineer.”
- “He is a brilliant engineer.”

Against attribute sentences like:

- “This job requires technical skill.”
- “This task involves nurturing children.”

By measuring similarities and conducting permutation-based significance testing, SEAT uncovers **contextual associations** that static embeddings would miss. It has proven effective in detecting subtle biases across deeper layers of transformer architectures.

4.3. ROME and Counterfactual Testing

The sense of using a more intervention-based approach, as opposed to working with biases passively, as is the case with WEAT and SEAT, introduces the Rank-One Model Editing (ROME) framework. ROME gives the researcher the ability to edit the memory of large language models (LLMs) like GPT-2 or GPT-3 directly, to test particular associations of knowledge in a cause and effect manner [23].

ROME, in practice, first locates and then alters internal representations, usually a key-value pair within the transformer in its attention mechanism to add or delete a particular association. As an instance, it would be possible to enforce the statement “The Eiffel Tower is in Berlin” into model and then witness the impact this has on associated outputs. ROME may be applied to:

- Insert biased facts (e.g., associating a profession with a gender)
- Remove harmful stereotypes
- Measure the downstream behavioral changes in text generation

This counterfactual capability enables a causal diagnosis of bias, rather than mere correlation. ROME thus opens the door to experimental debiasing and explainability strategies, offering insights into how bias is stored and can be surgically altered in LLMs.

4.4. Clustering and Causal Probing

Latent space clustering and causal probing are also another new trend in bias analysis. The rationale behind these approaches is that the biases may not manifest at individual vector level, but in the structural patterns that emerge in group level of the embedding space. Clustering methods detect the inherent vehicle grouping of embeddings (e.g. gendered names, racial identities) and compare their relative stances and orientations in the underlying space. These clusters are then applied in identifying biases by the comparative closeness of the cluster to features of being intelligent, aggressive or even moral.

Causal probing is even further but the idea is to see the degree to which an input feature (e.g., race or gender terms) and model output have a causal relationship. This is normally achieved:

- Modifying the input minimally (counterfactual generation)
- Measuring the change in model response (e.g., classification score or generated text)

- Controlling for confounding variables

In combination, these approaches provide a statistical, empirical insight into the effect of embedding structures on model behavior. These methods, unlike WEAT, that measures bias by surface-type associations, seek to reveal more in-depth biases in the learned representation at the causal pathways in the model.

5. Applications and Impact of WEAT

The consequences of introducing bias are more and more real as AI systems get more complex and impactful in society. The bias detected through the Word Embedding Association Test (WEAT) does not remain localized to academic examples; effects of the biases are far-reaching on the practices in the real world. Whether it is sentiment analysis engines and hiring algorithms, legal risk judgments, and healthcare diagnostics, the embedding layers that empower all these systems can also steer judgments in a segregated manner. The lens that is offered by WEAT is an important way of noticing these problems in advance; when they become ethical and legal, it is too late.

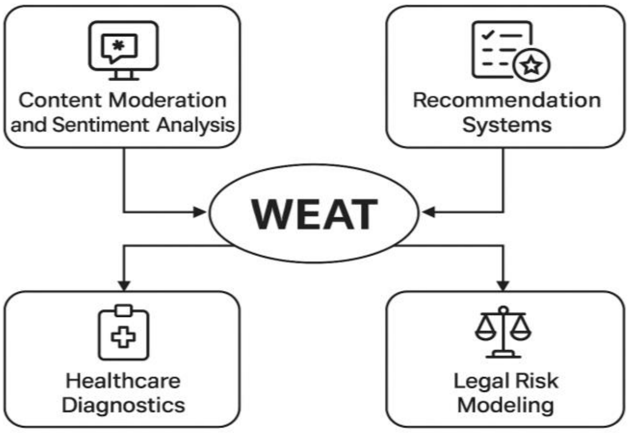

Fig 3: Applications of WEAT in Downstream NLP Tasks

5.1. Bias in Sentiment and Recommendation Systems

Sentiment analysis system auditing and recommendation engine auditing is one of the most straight-forward uses of WEAT. Pretrained embeddings are typically used to precondition such systems before testing on actual user input or output suggestions [24].

- In news recommendation, one-sided embeddings may correlate some demographic identifiers (e.g. religious or ethnic words) with negative sentiment, and may cause filtered or biased content delivery. Depending on information that is balanced or biased on such themes as immigration, policing, or health, a subtle embedding bias might determine whether a user will be shown balanced information or skewed.
- Word embeddings learned with biased corpora can cause biased scoring of candidates in the resume screening framework. In another case as far as the qualification is equal, words traditionally related to male dominated professions can be given more points than to their gender opposite. That is the idea that such implicit biases can be revealed by means of WEAT to analyze the strength of associations between terms of gender, on the one hand, and terms defining professional background, on the other hand.

With these quantitative associations becoming apparent, WEAT offers early-warning as a tool in the development of models to the developer and data scientists that may signal eventual ethical breach prior to models going into production.

5.2. Healthcare and Legal NLP Pipelines

Bias embedded by suggestion carries great stakes in the health and justice systems where the results of algorithms may have direct implications not only on human health but also their freedom. Healthcare NLP pipelines rely on models to generate insights (e.g. extract information) and/or predictions about clinical notes; or make recommendations (e.g. treatment recommendations). In case embedding links some types of diseases with sex, age, or even racial labels based on the tendencies in the past, it might give way to prejudiced diagnostics. As an example, it has been found that WEAT can be applied to identify situations in which the terms associated with the female have a lesser likelihood of being linked with high-risk situations and ultimately underdiagnosis may occur.

Within the law profession, some of the activities that the NLP can assist are recidivism forecasting, supervision risk estimation, and judicial document analysis. Systematic injustices here can be accentuated by encasing biases into them. By way of an example, any tight correlation between the names of specific ethnicities and any crime-related traits, however indirect, would defeat the risk scoring calculators in giving distorted scores. The models on which such disparities occur can be measured and documented through WEAT. These applications of use point to the value of WEAT as a diagnostic test but also as a compliance tool to ensure fair treatment involving high-impact decisions.

5.3. AI Governance and Ethical Oversight

Beyond technical applications, WEAT plays a growing role in shaping AI governance practices. As organizations adopt AI more broadly, there is an increasing demand for explainable, auditable, and ethical models especially in regulated sectors.

WEAT contributes to this effort by providing:

- Transparent documentation of embedding biases during model audits
- Baseline fairness evaluations for algorithm certification processes
- Quantitative support for human-in-the-loop decision-making systems

By adding the steps of dataset curation and model training to validation and deployment involving the WEAT to the AI development process, legal responsibility and liability can be mitigated and AI development can be held to any new reporting requirements within emerging AI regulations, such as the EU AI Act or the Algorithmic Accountability recently introduced in the U.S. Besides, the idea of bias testing tends to be proposed by ethical review boards and algorithmic impact assessments to be performed as due diligence. Embedding-level bias assessment by WEAT has increasingly become mandatory in practice and formed institutional norms and best practices.

6. Tools and Libraries Implementing WEAT

To make bias detection in word embeddings more accessible and standardized, several open-source tools and libraries have integrated support for WEAT and its variants. They represent so-called plug-and-play tool-kits, so that practitioners can use bias measures to many embedding models, datasets, and tasks without reinventing the wheel. This section points out three of the best libraries IBM AI Fairness 360[25], Word Embedding Fairness Evaluation (WEFE) framework, and Google Fairness Indicators that have distinct abilities in the audit of fairness in AI systems.

Table 3: WEAT-Capable Tools and Libraries

Tool/Library	WEAT Support	Model Compatibility	Supported Languages	Key Features
AI Fairness 360 (IBM)	Yes	Word2Vec, GloVe, BERT	English	Jupyter integration, SEAT support
WEFE	Yes	Any embedding model	Multilingual	Modular metrics, interactive reports
Fairness Indicators	Limited†	TensorFlow-based models	English	Visualization dashboards, TF-Extended APIs

Fairness Indicators do not natively implement WEAT but can visualize bias scores obtained via external testing.

6.1. AI Fairness 360 (IBM)

AI Fairness 360 (AIF360) is an inclusive open-source data bank advanced by IBM to assess and improve the biasness in machine learning models. It contains a wide variety of measures of fairness and debiasing algorithms, with native support to WEAT and SEAT. The AIF360 is built to support Python data classification and designed to fit smoothly with the Python-based data science workflow, particularly, in Jupyter Notebook settings. The library has ready to use implementations of WEAT which can enable the user to specify their own sets of target and attribute words words and administer the test using various embedding models. It also provides visualization of the effect sizes, p-values so that results can be interpreted by non-technical stakeholders as well. AIF360 allows the use of classic text embeddings (Word2Vec, GloVe) and can also use contextual embeddings (BERT) thereby making it more generalizable in NLP workflows.

6.2. WEFE (Word Embedding Fairness Evaluation)

WEFE is a pure Python package that aims at objectively computing the fairness of word embeddings. It provides a I/O set of modules using WEAT, SEAT and other association tests and it is highly configurable. Contrary to AIF360, which has a wider scope in ML, WEFE is designed to be embedding-centric with the possibility to compare experiments across languages and embedding models. The ability of WEFE to support cross-lingual testing of bias is another major strength and hence it can be employed wherever there is global application of NLPs. It gives the researcher the ability to specify the bias metrics programmatically or using JSON templates, and the output may be exported as the interactive HTML reports. The architecture of WEFE is so flexible that it can support pre-trained embeddings such as the fastText, gensim and even the custom-trained embedding models.

6.3. Fairness Indicators (Google)

Google has created Fairness Indicators a toolkit based on TensorFlow that focuses on fairness reporting and visualization in model evaluations. Originally, it has not built in WEAT, but can consume and view WEAT outputs alongside external testing frameworks. Practitioners operating in the TensorFlow Extended (TFX) ecosystem can benefit greatly in using the tool.

Fairness Indicators offers dashboard visuals of disparity statistics, including precision, recall, and false positive rates as separated into demographic segments. In cases where WEAT scores have already been computed in advance, these are able to be used added into the fairness analysis until the embedding level, to give a more in-depth picture of model behavior.

7. WEAT in Practice: Case Studies and Empirical Results

While the theoretical format of Word Embedding Association Test (WEAT) has long been developed, its actual effects can be realized to a greater extent when applied empirically. In the recent years: WEAT widely applied to various corpora, models and domains to bring out and measure biases. This part includes practical case studies in which WEAT has highlighted systemic problems that include old linguistic patterns to layer-by-layer analysis on transformer syntax and multilingual embeddings.

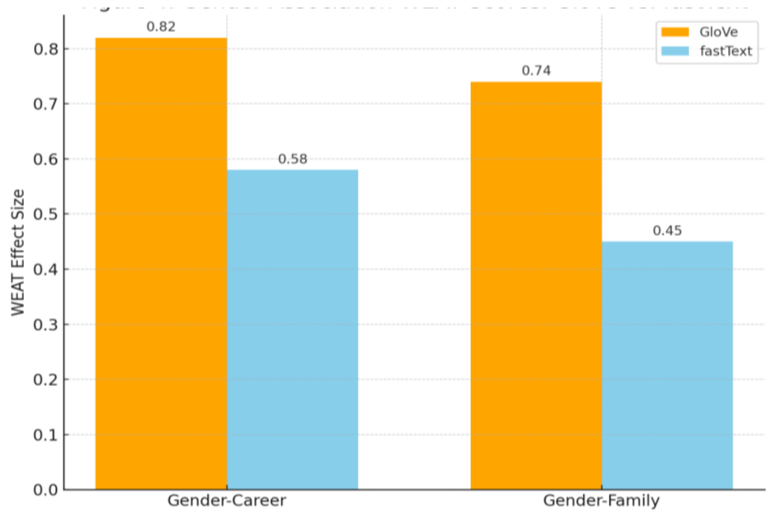

Fig 4: Gender Association WEAT Scores: GloVe vs. fastText

7.1. WEAT on Historical Corpora

Perhaps the most informed application of WEAT is that of temporal trends in language bias. Using WEAT to encode embeddings trained on historic-time collections like Google Books n-grams or historic-newspaper collections researchers have been able to quantify time-based societal change in attitudes. An example is a longitudinal study trained on the data between 1900 to 2000, where it shows that the connections between female and family to the male and career were strongest in the 1960s but the relationships were supposed to deteriorate in the 1990s. Other analogous experiments monitored racial and ethnic prejudice, e.g. the fluctuations in the correlation amid African-American names and negative value roots. Such trends are a reminder of the aspects of which WEAT is part of digital humanities and sociolinguistics as they aid in quantifying historical prejudice or advancement, providing an empirical look into the evolution of culture.

7.2. WEAT in Modern Pretrained Models

In modern NLP pipelines, static embeddings have largely been replaced by contextual language models like BERT, RoBERTa, GPT-2, and XLNet. To adapt WEAT for these architectures, researchers have performed layer-wise probing of token representations extracted from various depths of the model.

Findings indicate that:

- Lower layers in BERT tend to retain shallow syntactic structures, while middle layers encode the strongest biases.
- Later layers, though more semantically abstract, still preserve some biased associations, especially in models like GPT-2 trained on unfiltered web data.

When applied to sentence-level templates (as in SEAT), WEAT reveals that gender and race-related biases persist across multiple transformer variants. GPT-based models, in particular, have shown susceptibility to amplifying stereotypes in generation tasks when prompted with biased seed phrases. These results emphasize the need for layer-specific mitigation strategies and deeper transparency in model auditing.

7.3. Cross-Lingual and Domain-Specific WEAT

Recent studies have extended WEAT to multilingual and domain-specific embeddings. In cross-lingual settings, WEAT has been used to compare gender biases across languages such as English, Spanish, German, and Hindi. Notably, gendered languages (e.g., Spanish, German) often exhibit stronger stereotypical associations than more neutral languages.

In domain-specific applications, embeddings trained on:

- Clinical notes have shown gender bias in disease attribution (e.g., associating “anxiety” more with female terms),
- Legal documents have revealed racial correlations with terms like “guilty” or “risk,” especially in predictive policing datasets.

These results show that WEAT is not limited to general-purpose NLP models it can be repurposed for specialized contexts, making it highly relevant in high-stakes industries where equity and compliance are critical.

8. Challenges and Future Directions

However, the Word Embedding Association Test (WEAT) is not perfect in spite of its popularity and influence.A set of unresolved problems threatens the validity of the WEAT method of detecting bias and its universality and interpretability. These constraints are caused both by technical factors and more general questions of the epistemology of bias in general. This part irons out some of the primary open issues and discusses the multidisciplinary avenues that the development of WEAT and its descendants must travel along to make any significant progress.

Table 4: Summary of Unresolved Challenges in Bias Detection via WEAT

Challenge Area	Description	Consequences
Definitional Ambiguity	Bias lacks a consistent definition across cultures and disciplines	Limits transferability of WEAT across domains or languages
Contextuality in Models	Transformer models generate dynamic meanings that evade static testing	Reduces WEAT’s effectiveness on modern NLP architectures
Subjectivity of Word Lists	Outcomes depend on manually curated target/attribute terms	Risk of researcher bias or reproducibility issues
Limited Causal Interpretability	WEAT measures associations, not causation	Difficult to determine real-world behavioral impacts
Lack of Legal Standardization	No regulatory consensus on audit thresholds	Hinders formal use in compliance or governance frameworks

8.1. Definitional Ambiguities of Bias

One of the most fundamental challenges in bias detection is the lack of a universally accepted definition of “bias.” What is considered biased in one sociocultural context may be neutral or even desirable in another. This issue of cultural relativism complicates the creation of globally applicable bias detection benchmarks. Furthermore, bias can be framed in either descriptive (what the model encodes) or normative (what the model should encode) terms. WEAT, as a statistical tool, is primarily descriptive it tells us whether associations exist, but not whether those associations are ethically problematic. This distinction is crucial when models are deployed in sensitive or regulated environments. Without clear normative guidelines, it becomes difficult to interpret WEAT scores beyond comparative studies, which limits its prescriptive utility in policymaking or compliance.

8.2. Model Agnosticism vs. Contextuality

WEAT was originally designed for static embedding models, making it relatively model-agnostic in early NLP ecosystems. However, with the dominance of contextual transformer models, this strength has become a weakness. In transformers, the meaning of a word changes dynamically with its context, making traditional vector-based comparisons insufficient. Although extensions like SEAT have attempted to bridge this gap, their effectiveness varies across tasks, architectures, and layers. For instance, a word may display bias in the middle layers of BERT but not in the output layer. This flexibility begs the question of how and where to measure bias in complex large and opaque models. Besides, the growing popularity of prompt- or fine-tuned models introduces another level of complication because even minor changes to training task or prompt may influence biased behaviors. Accordingly, researchers stand at a cross-purpose between the ease of use of WEAT and the semantics fluidity of contemporary models.

8.3. Need for Multidisciplinary Integration

To advance beyond current limitations, bias detection must embrace interdisciplinary collaboration. At its core, WEAT draws inspiration from psychometrics, but its full potential can only be realized when integrated with legal, ethical, and social frameworks:

- In law, understanding how embedding bias influences sentencing decisions requires mapping statistical associations to legal standards of fairness.
- In healthcare, bias must be evaluated not only by model metrics but also by clinical outcome disparities across demographic groups.
- In public policy, embedding audits could be part of algorithmic impact assessments, similar to environmental or financial audits.

- Some possible future versions of WEAT might include adaptive benchmarks based on community feedback, causal test processes, or even policy-driven goals at debiasing. Crossing the boundary between computational and social sciences will be key to building powerful, context-sensitive standards of fairness that will be technically valid and socially relevant.

9. Conclusion

As artificial intelligence is infiltrating decision-making in multiple instances of vital activities, the need to enforce equity and responsibility through language models has become an urgent necessity. Among the most influential and user-friendly tests of diagnosing word vegetative biases, one should note the Word Embedding Association Test (WEAT). Since it incorporates principles developed by cognitive psychology into a computational format, WEAT allows researchers and practitioners to measure how extensively social stereotypes are coded to AI systems. This review has followed history of the development of WEAT starting with its abstract and theoretical deductions continuing to mathematical bases and its practical implementation. We explained how WEAT has been generalized using such tools as SEAT and ROME to cover more dynamic, modern transformer models, and how it has been applied to practical applications such as medicine, employment, news promotion and legal analysis. The integration of WEAT in these libraries such as AI Fairness 360 and WEFE has only increased the level of democratization of embedding-level audits, which is critical to recommend its use. Although all this has been developed, the test also has its problems. WEAT applies certain assumptions regarding language use, definition of bias and statistical significance which may not be valid in other contexts (cultural or technical).

Moreover, its weaknesses related to the ability to measure the contextual and causal aspects of bias indicate the necessity of innovative methodology. With the shift to the age of large, cross-lingual and context-aware language models, the community should not only leave behind association-based metrics but also shift towards more comprehensive, explainable, and causally informed evaluation principles. Looking ahead, the future of bias detection in AI will rely on deeper interdisciplinary collaboration combining insights from computer science, linguistics, sociology, law, and ethics. Policymakers and developers alike must recognize that bias is not solely a technical artifact, but a reflection of the data, norms, and values embedded in our digital systems. WEAT has served as a catalyst in this journey, but it should now be seen as a foundation upon which more advanced and accountable auditing tools can be built. WEAT remains a cornerstone of AI fairness research. It provides a vital entry point for evaluating bias in embeddings and helps pave the way for ethically responsible and socially aligned AI systems. Continued refinement, contextual adaptation, and policy integration will ensure that WEAT and the principles behind it remain central to the responsible development and deployment of AI.

References

[1] M. Kuziemski and G. Misuraca, "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings," Telecommunications Policy, vol. 44, no. 6, p. 101976, 2020.

[2] J. Whittlestone et al., "Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research," Nuffield Foundation, London, 2019, pp. 1–59

[3] R. Patil et al., "A survey of text representation and embedding techniques in NLP," IEEE Access, vol. 11, pp. 36120–36146, 2023.

[4] T. Bolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in Advances in Neural Information Processing Systems, vol. 29, 2016.

[5] K. Ethayarajh, D. Duvenaud, and G. Hirst, "Understanding undesirable word embedding associations," arXiv preprint arXiv:1908.06361, 2019.

[6] J. Jangid and S. Dixit, The AI Renaissance: Innovations, Ethics, and the Future of Intelligent Systems, vol. 1, Technoscience Academy, 2023.

[7] F. Ahmed, "Cloud Security Posture Management (CSPM): Automating Security Policy Enforcement in Cloud Environments," ESP International Journal of Advancements in Computational Technology (ESP-IJACT), vol. 1, no. 3, pp. 157–166, 2023.

[8] O. Papakyriakopoulos et al., "Bias in word embeddings," in Proc. 2020 Conf. on Fairness, Accountability, and Transparency (FAT), 2020.

[9] P. P. Liang et al., "Towards understanding and mitigating social biases in language models," in Proc. Int. Conf. Machine Learning (ICML), PMLR, 2021.

[10] A. Van Loon and J. Freese, "Word embeddings reveal how fundamental sentiments structure natural language," American Behavioral Scientist, vol. 67, no. 2, pp. 175–200, 2023.

[11] K. Durrheim et al., "Using word embeddings to investigate cultural biases," British Journal of Social Psychology, vol. 62, no. 1, pp. 617–629, 2023.

[12] K. K. Singh et al., "Don't judge an object by its context: Learning to overcome contextual bias," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[13] R. Xu et al., "Word embedding composition for data imbalances in sentiment and emotion classification," Cognitive Computation, vol. 7, pp. 226–240, 2015.

[14] S. Dixit, "AI-powered risk modeling in quantum finance: Redefining enterprise decision systems," Int. J. Scientific Research in Science, Engineering and Technology, vol. 9, no. 4, pp. 547–572, 2022. doi: 10.32628/IJSRSET221656

[15] F. Yashu et al., "Thread mitigation in cloud native application development," Webology, vol. 18, no. 6, pp. 10160–10161, 2021. [Online]. Available: https://www.webology.org/abstract.php?id=5338s

[16] N. Swinger et al., "What are the biases in my word embedding?," in Proc. 2019 AAAI/ACM Conf. on AI, Ethics, and Society, 2019.

[17] 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, p. e0231189, 2020.

[18] F. Ahmed, "Quantum-Resistant Cryptography for National Security: A Policy and Implementation Roadmap," Int. J. Multidisciplinary on Science and Management, vol. 1, no. 4, pp. 54–65, 2024.

[19] U. Naseem et al., "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," Trans. Asian and Low-Resource Language Information Processing, vol. 20, no. 5, pp. 1–35, 2021.

[20] S. Husse and A. Spitz, "Mind your bias: A critical review of bias detection methods for contextual language models," arXiv preprint arXiv:2211.08461, 2022.

[21] J. Jangid, "Efficient Training Data Caching for Deep Learning in Edge Computing Networks," Int. J. Scientific Research in Computer Science, Engineering and Information Technology, vol. 7, no. 5, pp. 337–362, 2020. doi: 10.32628/CSEIT20631113

[22] C. May et al., "On measuring social biases in sentence encoders," arXiv preprint arXiv:1903.10561, 2019.

[23] J. M. Alvarez and S. Ruggieri, "Counterfactual situation testing: Uncovering discrimination under fairness given the difference," in Proc. 3rd ACM Conf. on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2023.

[24] I. Rabiu et al., "Modeling sentimental bias and temporal dynamics for adaptive deep recommendation system," Expert Systems with Applications, vol. 191, p. 116262, 2022.

[25] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.