DEVELOPING A METRIC TO DETECT GENDER BIAS IN CONTEXTUAL WORD EMBEDDINGS.

H.M.I.K. Dhanawardhana

IT21183690

B.Sc. (Hons) Degree in Information Technology Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

April 2024

DEVELOPING A METRIC TO DETECT GENDER BIAS IN CONTEXTUAL WORD EMBEDDINGS.

H.M.I.K. Dhanawardhana.

IT21183690

Dissertation submitted in partial fulfilment of the requirements for the Special Honours Degree of Bachelor of Science in Information Technology Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

April 2024

DECLARATION

I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 04/11/2025

The above candidate has carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the Supervisor : Date: 04/11/2025

ACKNOWLEDGEMENT

This research would not have been possible without the invaluable support, guidance, and encouragement of many individuals and organizations, to whom I am profoundly grateful.

Heartfelt gratitude is extended to my supervisor, Dr. Prasanna S. Haddela, for his immense support and guidance throughout this journey. His invaluable feedback, unwavering encouragement, and commitment made him an exceptional mentor, ensuring the successful completion of this research.

Sincere thanks are extended to my co-supervisor, Ms. Thisara Shyamalee, for her valuable input and encouragement, which greatly contributed to refining and enhancing this work.

Their contributions in making avian data accessible were crucial to the foundation of this research.

Finally, heartfelt thanks are extended to Sri Lanka Institute of Information Technology (SLIIT) for the resources, facilities, and academic environment that made this work possible.

To all those who supported me directly or indirectly, your contributions have been profoundly meaningful, and I am deeply grateful.

ABSTRACT

As transformer-based language models increasingly underpin decision-making in high-stakes domains, detecting and mitigating gender bias in contextual word embeddings has become a critical challenge in responsible AI. This study introduces the Context-Aware Bias Metric (CABM) a novel, interpretable, and multi-dimensional framework for quantifying gender bias in transformer-based embeddings such as BERT. CABM integrates three core features: cosine similarity (semantic bias), Pointwise Mutual Information (statistical co-occurrence bias), and sentence-level contextual shift (embedding variability across gendered contexts). Using the WinoBias dataset (927 annotated sentences), embeddings were extracted for occupation terms in both male- and female-framed sentence structures.

Bias scores were computed through three weighting strategies PCA, PCA + Random Forest, and SHAP + Random Forest to balance robustness and interpretability. Statistical analyses including Mann-Whitney U, Shapiro-Wilk, and correlation tests confirmed CABM’s validity and discriminative power. Among the strategies, SHAP + RF emerged as the most stable and interpretable method, achieving 69.2% alignment with gendered pronoun predictions in a masked language model, and demonstrating strong internal consistency (Pearson r = 0.93 with PCA + RF). Visualization techniques such as SHAP bar plots and KDE distributions illuminated occupation-specific bias patterns and emphasized the dominant role of contextual embedding shifts.

This work advances fairness diagnostics by offering a modular, extensible, and explainable approach to bias quantification in NLP, with demonstrated generalizability across embedding behaviors and practical alignment with real model outputs.

Keywords: Contextual Word Embeddings, Gender Bias Detection, BERT, SHAP, NLP Fairness, Context-Aware Metric, Transformer Models, Cosine Similarity, PMI, Sentence-Level Bias, Bias Evaluation, CABM, Explainable AI

Table of Contents

LIST OF FIGURES

LIST OF ABBRIVIATIONS

Abbreviation	**Description**
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CABM	Context-Aware Bias Metric
CI/CD	Continuous Integration / Continuous Deployment
CLS	Classification Token (used in BERT)
CNN	Convolutional Neural Network
KDE	Kernel Density Estimation
LLM	Large Language Model
MAC	Mean Average Cosine Similarity
ML	Machine Learning
MLM	Masked Language Modeling
NLP	Natural Language Processing
PCA	Principal Component Analysis
PMI	Pointwise Mutual Information
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
UBM	Unified Bias Metric (used in literature comparisons)
WEAT	Word Embedding Association Test

Abbreviation	**Description**
WEFE	Word Embedding Fairness Evaluation

1. INTRODUCTION

1.1 Background & Literature survey

1.1.1The Evolution of Word Embeddings in NLP

Natural Language Processing (NLP) has evolved significantly over the past two decades, largely driven by innovations in how words and phrases are numerically represented. The foundation of most modern NLP systems lies in **word embeddings**, which map words to dense vector spaces where semantic similarity is preserved. These embeddings allow algorithms to process language in a more meaningful way by capturing word relationships through mathematical proximity.

Early word embedding models such as **Word2Vec** and **GloVe** revolutionized NLP by representing words as fixed-length vectors based on their co-occurrence patterns in large corpora. **Word2Vec**, introduced by Mikolov et al. [1], relies on predicting surrounding words (Skip-Gram) or predicting the target word from its context (CBOW), creating embeddings that encode syntactic and semantic regularities. **GloVe**, developed by Pennington et al. [2], takes a matrix factorization approach, utilizing global co-occurrence statistics to generate similar vector spaces. These embeddings enabled analogical reasoning (e.g., *king - man + woman ≈ queen*) and led to rapid performance improvements in tasks such as sentiment analysis, part-of-speech tagging, and named entity recognition.

However, one major limitation of these early models is their **static nature**: each word has a single embedding regardless of context. This causes ambiguity in words with multiple meanings (e.g., *"bank"* in *"river bank"* vs. *"investment bank"*) and fails to capture subtle shifts in meaning depending on usage. To address this, the field moved toward **contextual embeddings**, marking a paradigm shift in NLP.

Contextual word embeddings, introduced by models such as **ELMo** and later advanced by **BERT (Bidirectional Encoder Representations from Transformers)** [3], generate dynamic representations based on the entire sentence context. Instead of assigning one embedding per word, these models compute token-level embeddings that vary depending on surrounding words. For example, BERT will produce a different vector for the word *"doctor"* in the sentence *"She is a doctor"* compared to *"He is a doctor."* This breakthrough allowed models to better handle polysemy, syntax, and discourse structure.

Pretrained on massive corpora using masked language modeling and next sentence prediction tasks, **BERT** and its successors (e.g., RoBERTa, XLNet, ALBERT) have become the backbone of state-of-the-art systems in tasks such as question answering, coreference resolution, and text classification. However, despite their increased performance, these models are not immune to societal bias. Since contextual embeddings are derived from human-generated text data, they often reflect implicit stereotypes and cultural biases embedded in the training corpora.

As **Caliskan et al.** [4] revealed, even models trained on neutral-sounding data can encode deep-seated human biases, such as associating male terms with careers and female terms with family. This phenomenon, which first emerged in static embeddings, has now been shown to extend to contextual representations as well, albeit in more complex and subtle ways.

The emergence of contextual embeddings has thus made traditional bias detection techniques insufficient. Fixed-word association methods, like WEAT, struggle to capture biases that arise due to sentence-level constructions, co-reference patterns, or interaction between words. For example, a bias might not be detectable at the word level (e.g., "doctor") but becomes apparent when analyzed in context (e.g., “She is a doctor” being predicted differently than “He is a doctor”).

This shift underscores the urgent need for new bias detection methodologies that go beyond static representations and account for contextual dependencies. As transformer-based models become increasingly integrated into real-world applications including hiring platforms, healthcare systems, and social media moderation tools there is a growing responsibility to understand and mitigate the embedded biases they may carry.

1.1.2How Bias Emerges in Language Models

While advancements in word embeddings have transformed NLP systems, they have also introduced serious ethical concerns most notably, the encoding and perpetuation of social biases. Language models trained on large-scale web data, news articles, books, and online forums inevitably learn from historical and cultural patterns present in those texts. Consequently, these models replicate and reinforce stereotypes, even when such associations are undesirable or harmful.

Bias in language models typically manifests in three forms:

1. **Occupational Stereotyping**: Certain professions (e.g., "nurse," "engineer," "teacher") are disproportionately associated with one gender. Studies such as Caliskan et al. [4] and Bolukbasi et al. [5] demonstrate how embeddings align male terms with technical fields and leadership roles, while associating female terms with domestic or support roles.
2. **Syntactic Bias in Coreference**: In coreference resolution tasks, pronouns like *"she"* and *"he"* are more likely to be linked to gender-stereotyped occupations, even when the context is ambiguous. Zhao et al. [6] exposed this issue using the WinoBias dataset, showing that models favored male associations in neutral contexts, revealing deeply ingrained biases.
3. **Semantic Association and Polarity**: Beyond simple co-occurrence, embeddings capture sentiment and moral association. For example, words like “leader” or “intelligent” may appear closer to male pronouns, while “nurturing” or “emotional” may skew toward female pronouns [7], [8].

These biases are not just theoretical; they have been observed in real-world deployments. AI-powered résumé screening tools have favored male candidates for leadership roles. Chatbots have echoed sexist stereotypes. Even in language generation, models complete prompts like *“The doctor said that…”* with male pronouns more frequently than female ones.

A key factor contributing to these biases is the distributional hypothesis underpinning most NLP systems: *“You shall know a word by the company it keeps.”* While useful for semantic modeling, this principle can lead to misaligned associations when societal usage patterns are skewed. For example, if online corpora disproportionately describe women in caregiving contexts, the model learns and amplifies that framing.

Furthermore, as Manzini et al. [9] and Gonen & Goldberg [8] note, debiasing efforts often fall short. Techniques like projection-based debiasing may reduce bias in certain evaluations, but latent gender subspaces remain, allowing downstream models to still detect and exploit bias-encoded patterns. These “bias remnants” are especially prevalent in contextual embeddings, where token representations are heavily influenced by surrounding text and subtle phrase structures.

As transformer-based models like BERT [3] and GPT are fine-tuned on domain-specific tasks, these biases may become more pronounced. For instance, fine-tuning on medical or legal text may amplify pre-existing gender associations in those domains. The issue becomes even more complex when embeddings are used in multi-modal systems (e.g., CLIP), where language and image biases can reinforce each other.

In summary, bias in language models is not an anomaly it is an emergent property of data-driven learning. It stems from real-world inequalities, encoded and preserved through distributional semantics. Without proper evaluation and mitigation, such bias has the potential to entrench and amplify social injustice. This motivates the need for more context-sensitive, explainable, and robust methods to quantify and address gender bias in NLP systems, especially those relying on contextual embeddings.

1.1.3Existing Bias Detection Metrics

As the field of NLP matured and the social implications of model predictions became more evident, researchers began developing metrics to measure bias in word embeddings. These metrics aim to detect and quantify the extent to which word vectors encode discriminatory or stereotypical associations. While several prominent methods exist, most are designed for static embeddings, and many fall short when applied to contextual language models like BERT

Word Embedding Association Test (WEAT)

The WEAT metric, introduced by Caliskan et al. [4], is among the most well-known tools for measuring bias in word embeddings. Inspired by the Implicit Association Test in psychology, WEAT assesses the association strength between target groups (e.g., “man,” “woman”) and attribute groups (e.g., “career,” “family”) based on cosine similarity. A significantly higher similarity between male terms and career-related words, compared to female terms, indicates gender bias.

While WEAT helped uncover major issues in static embeddings, it has two main limitations: (1) it operates on word sets, meaning it does not capture token-level or sentence-level behavior, and (2) it assumes that bias is symmetrically measurable, which is often not the case in real-world corpora.

Mean Average Cosine Similarity (MAC)

MAC extends the idea of cosine similarity by computing the **average cosine distance** between gendered terms and context terms across the embedding space. This metric is simpler and more interpretable than WEAT but shares similar drawbacks. It does not account for **frequency-based bias** and struggles with **contextual nuances** present in modern transformers.

Pointwise Mutual Information (PMI)

PMI, originally a statistical metric, has been adapted to measure gender bias in word co-occurrence. For example, Valentini et al. [10] use PMI to examine how often gendered pronouns (“he,” “she”) co-occur with certain occupations in large corpora. Higher PMI values indicate stronger associations. Although PMI provides useful frequency-based insights, it fails to account for semantic meaning, which is crucial in modern NLP systems. Moreover, PMI scores can become unstable when co-occurrence counts are low, especially in rare or balanced datasets.

SAME Score and StereoSet

The **SAME score**, introduced in the **StereoSet benchmark**, evaluates whether a model prefers stereotypical completions in context. For example, it compares model preferences between sentences like:

- “The nurse helped the patient because she was kind.” (stereotypical)
- “The nurse helped the patient because he was kind.” (anti-stereotypical)

While useful, SAME relies on **template-based sentence construction**, limiting generalizability. It also evaluates bias **based on sentence likelihood**, which is model-dependent and may not reflect real semantic encoding.

WEFE: Word Embedding Fairness Evaluation

To bring consistency to bias evaluation, Badilla et al. [11] introduced **WEFE**, a unified framework that provides a common API and dataset for running fairness metrics across embedding types. WEFE includes tools to measure direct bias, association bias, and even differential bias across languages. However, like other tools, WEFE still primarily operates at the **token or word-set level**, lacking the **sentence-level context analysis** necessary for transformer-based embeddings.

Metric	Core Technique	Strengths	Key Limitations
WEAT	Cosine similarity between predefined word groups	Simple, intuitive, popular	Fails to account for sentence-level context; assumes static word meaning
MAC	Average cosine similarity	Easy to compute; interpretable	Ignores co-occurrence frequency and syntactic variations
PMI	Statistical co-occurrence of gendered terms and occupations	Captures frequency-based associations	Unstable with low-frequency data; does not capture semantic or contextual meaning
SAME (StereoSet)	Likelihood-based preference for stereotypical vs. anti-stereotypical completions	Operates on full sentences	Relies on fixed templates; lacks generalization across real-world usage
WEFE	Unified evaluation framework for embedding bias metrics	Supports multiple fairness metrics	Still focuses on word-level association; lacks interpretability and context sensitivity

1.1.5The Need for Contextual Bias Detection

As transformer-based models like BERT, RoBERTa, and GPT have become foundational in natural language processing, the structure and behavior of word embeddings have evolved significantly. Unlike static embeddings, which assign a single vector to each word, **contextual word embeddings** generate dynamic, token-specific vectors that depend on the surrounding sentence. This shift has fundamentally altered how bias is encoded and manifested in language models **moving from word-level to context-level**.

Traditional bias metrics, such as WEAT, MAC, and PMI, are **ill-suited to capture the complexities of these modern embeddings**. These metrics operate on the assumption that word meaning is fixed, ignoring how the presence of pronouns, sentence structure, or syntactic roles can alter the meaning and therefore, the bias of a word.

Consider the word *“engineer.”* In isolation, this term may appear gender neutral. However, when evaluated in context e.g., *“She is an engineer”* versus *“He is an engineer”* BERT may produce different embeddings for the same word. This shift is not detectable by static metrics. Similarly, coreference resolution systems may treat *"The doctor said he was tired"* differently than *"The doctor said she was tired,"* even when the sentences are semantically identical. These subtle variations reflect contextual gender bias, a phenomenon that cannot be captured using word lists or cosine distances alone.

Moreover, bias can now emerge at the sentence level, where the embedding of an entire sentence changes based on the gendered pronouns used, the role of the speaker, or the order of clauses. These changes are especially problematic in applications such as:

- Conversational agents, where sentence framing affects responses
- Resume filtering systems, where sentence construction may bias evaluation
- Medical diagnosis systems, where gendered references could skew predictions

As noted by Guo and Caliskan [10], transformer-based embeddings encode intersectional biases, meaning that words or phrases can carry different bias profiles based on their full context an issue nearly impossible to detect using one-dimensional bias scores. Similarly, Gonen and Goldberg [8] warn that even after applying standard debiasing methods, residual biases persist in the contextual behavior of words.

This growing complexity necessitates a new class of bias detection tools that:

- Operate at the sentence level
- Capture contextual embedding shifts
- Integrate multiple signals (semantic, statistical, and contextual)
- Remain interpretable and explainable

To address these needs, this research proposes the Context-Aware Bias Metric (CABM) a unified metric that combines cosine similarity, PMI, and sentence-level contextual bias, with weights derived using robust techniques such as PCA, Random Forest, and SHAP. By doing so, CABM offers a more holistic, adaptable, and transparent framework for measuring gender bias in contextual embeddings.

1.1.6Literature Insights and Justification for CABM

Across the past decade, growing awareness of gender bias in machine learning has led to the development of several evaluation frameworks, bias metrics, and debiasing strategies. These contributions have laid a strong foundation, but they also reveal a consistent pattern: existing metrics are either too narrow, too simplistic, or not designed for contextual language models.

For instance, early work by Caliskan et al. [4] introduced the Word Embedding Association Test (WEAT), which showed that even embeddings trained on neutral data encode human-like stereotypes. This exposed the need for automated tools to audit bias in NLP models. Following this, Bolukbasi et al. [5] proposed a debiasing algorithm that identifies a gender subspace in word embeddings and neutralizes words that should be gender-independent. While effective for static vectors, this method cannot be directly applied to contextual embeddings, where each token's representation changes dynamically depending on its usage.

Later, researchers explored co-occurrence-based metrics, such as PMI. For example, Valentini et al. [11] demonstrated that PMI-based gender bias metrics suffer from frequency instability, particularly in balanced or limited datasets. Their findings revealed that high PMI values can emerge even when bias is not strongly encoded semantically, making the metric unreliable on its own.

To improve reliability, Badilla et al. [12] introduced the WEFE framework, which standardizes evaluation across multiple bias metrics. However, WEFE still relies on static measures like cosine similarity and lacks the ability to analyze how contextual embeddings behave in full sentences. Moreover, metrics like SAME [13] and MAC, while useful in capturing preference or association, do not explain why a bias score is high—limiting their value in interpretability and accountability.

Other researchers have focused on debiasing contextual models directly. Zhao et al. [6], for example, used the WinoBias dataset to reveal gender bias in coreference systems and proposed data augmentation as a mitigation technique. While effective, this approach addresses the model behavior, not the bias quantification itself. More recently, Guo and Caliskan [10] revealed that even intersectional and emergent biases can be encoded in BERT-style embeddings, further highlighting the need for sentence-level evaluation.

Despite these developments, there remains no unified, interpretable, and context-aware metric that integrates:

- Semantic similarity (e.g., cosine bias)
- Statistical bias (e.g., PMI)
- Sentence-level embedding variation (e.g., contextual shift in BERT outputs)

This research addresses this gap by proposing the Context-Aware Bias Metric (CABM) a novel framework that fuses multiple dimensions of bias into a single, interpretable score. CABM computes semantic, statistical, and contextual bias features for occupation-related terms and integrates them using data-driven weighting strategies such as:

- Principal Component Analysis (PCA)

- PCA + Random Forest
- SHAP + Random Forest

These methods ensure that CABM is not only robust across different linguistic structures but also transparent providing insight into how each component contributes to the overall bias score. The metric is validated using the WinoBias dataset, which contains both semantic and syntactic sentence pairs, making it ideal for evaluating contextual models like BERT.

By addressing the weaknesses of prior methods and aligning with current challenges in transformer-based NLP systems, CABM presents a significant advancement in the measurement of gender bias in modern language models.

1.2 Research Gap

Despite the proliferation of bias detection methods in NLP, most existing approaches fall short when applied to contextual word embeddings—particularly at the sentence level, where transformer-based models like BERT operate most effectively. These limitations undermine the reliability and interpretability of current bias evaluation strategies, especially in real-world applications where decisions are made based on full sentences or documents rather than isolated words.

One of the most pressing issues is the overreliance on word-level representations, which do not capture how meaning and bias shift dynamically depending on surrounding words. Traditional metrics such as WEAT and MAC focus on individual words and static embeddings, which are no longer representative of how modern models process language [4], [13]. In contextual models, the same word can have significantly different representations across various sentences, making token-level evaluation insufficient for detecting subtle or situational bias.

Another key limitation is the use of frequency-sensitive metrics, such as cosine similarity and Pointwise Mutual Information (PMI), which often become unreliable in low-frequency or sparse data conditions. Valentini et al. [11] highlight how PMI-based scores can misrepresent associations due to uneven word distributions, especially in specialized domains or small corpora. These instabilities can lead to both false positives and missed biases, reducing the generalizability of results.

Additionally, most current approaches fail to unify semantic, statistical, and contextual signals into a cohesive metric. Bias is multi-dimensional: it involves not just proximity in embedding space (semantic similarity), but also statistical co-occurrence patterns and contextual shifts in meaning. By treating these signals in isolation, existing tools lose explanatory power and often miss deeper, systemic patterns of bias.

Interpretability is another major challenge. As Badilla et al. [12] point out in their critique of the

WEFE framework, many metrics provide raw scores without insight into what specific linguistic or contextual features contribute to the measured bias. This lack of transparency makes it difficult for researchers and practitioners to diagnose or mitigate bias effectively in real-world systems.

Moreover, most bias detection tools focus narrowly on gender bias, often using binary male/female associations. While this is a critical area of study, it neglects intersectional biases the overlapping and compounding effects of multiple identity dimensions like race, gender, and class. Guo and Caliskan [10] demonstrate that contextual embeddings reflect these complex interactions, which require richer and more flexible tools for detection and analysis.

Collectively, these gaps reveal the need for a comprehensive, interpretable, and context-sensitive bias detection metric one that can operate at the sentence level, integrate semantic similarity, statistical co-occurrence, and contextual embedding variation, adapt to modern transformer architectures like BERT, and provide explainable outputs that reveal the sources and patterns of bias.

1.3 Research Problem

Building on the limitations identified in existing bias detection methods, this research addresses the following central problem:

How can gender bias in contextual word embeddings be effectively measured by integrating semantic similarity, statistical co-occurrence, and sentence-level embedding behavior into a unified, interpretable metric?

This problem emphasizes not only the detection of bias but also the development of a composite, multi-dimensional score that captures bias from different perspectives semantic alignment, corpus-based association patterns, and contextual variability. Unlike traditional approaches that operate at the word level or rely on a single measurement signal, this research seeks to formulate a metric that can quantify bias holistically in contextual embeddings produced by models like BERT. The ultimate goal is to produce a tool that is both statistically rigorous and practically interpretable, bridging the gap between theoretical fairness analysis and real-world NLP evaluation.

1.4 Research Objectives

To address the challenges outlined in the research problem, this study establishes a set of focused objectives aimed at designing, implementing, and validating a novel metric for detecting contextual gender bias in transformer-based language models. The overarching goal is to develop an interpretable, data-driven framework that integrates semantic, statistical, and sentence-level context into a unified bias evaluation strategy suitable for contextual word embeddings.

Specific Objectives:

- To define and formalize contextual gender bias in the domain of word embeddings, with a particular focus on how sentence structure, pronoun usage, and model context influence gender associations beyond what traditional metrics capture.
- To propose and implement a unified metric Context-Aware Bias Metric (CABM) capable of quantifying gender bias in contextual embeddings by:
    - Calculating Cosine Similarity between gendered pronouns and occupational terms to capture semantic associations.
    - Computing Pointwise Mutual Information (PMI) to evaluate statistical co-occurrence bias between gendered terms and professions.
    - Measuring Sentence-Level Contextual Bias by analyzing embedding shifts of occupation terms across male- and female-gendered sentence frames using BERT.
- To apply and compare multiple feature-weighting approaches for optimizing the CABM formula and ensuring interpretability:
    - Principal Component Analysis (PCA)
    - PCA combined with Random Forest
    - SHAP (SHapley Additive Explanations) combined with Random Forest
- To validate the CABM metric using the WinoBias dataset, which contains syntactically and semantically controlled sentence pairs designed to test stereotypical and anti-stereotypical gender roles.
- To perform statistical validation of CABM across different bias components using:
    - Mann-Whitney U Test for distributional bias
    - Shapiro-Wilk Test for normality
    - Chi-Square Test for categorical association
- To compare CABM with existing bias metrics (e.g., WEAT, PMI, MAC, SAME) and

highlight its effectiveness in capturing subtle, context-sensitive biases that remain undetected by traditional methods.

- To enhance explainability and transparency through:
  - SHAP-based feature attribution visualizations
  - Contextual embedding comparison plots
  - Heatmaps and bar graphs illustrating gender bias trends across occupations and sentence types
- To ensure reproducibility and extensibility of the proposed framework by providing open-source implementation, enabling future adaptation to other bias domains (e.g., racial or age bias) and other contextual models (e.g., RoBERTa, GPT).

2. METHODOLOGY

2.1 Introduction to Methodology

This chapter details the methodological framework developed to effectively measure and analyze gender bias within contextual word embeddings. Building on the identified research gaps and objectives outlined in the Introduction chapter, this methodology systematically integrates multiple computational dimensions into a unified metric, thereby addressing the limitations present in existing bias detection approaches.

The primary goal of this research is to develop a novel, comprehensive, and interpretable metric Context-Aware Gender Bias Score (CABM) capable of accurately detecting gender bias in contextual embeddings, specifically those generated by transformer-based models such as BERT. This metric is designed to overcome significant drawbacks found in traditional bias measurement techniques, particularly their inability to capture nuanced, sentence-level contextual biases effectively. To achieve this overarching objective, the research specifically pursues the following detailed sub-objectives:

The methodology begins by extracting these critical bias-related features from the widely recognized WinoBias dataset, which includes carefully structured sentences designed to test both stereotypical and anti-stereotypical associations in semantic and syntactic contexts. Following extraction, the research employs three advanced statistical and machine-learning-based weighting approaches

Principal Component Analysis (PCA), PCA combined with Random Forest Feature Importance, and Random Forest combined with SHapley Additive exPlanations (SHAP) to optimally integrate these features into the CABM. PCA provides interpretability through dimensionality reduction, PCA combined with Random Forest captures non-linear interactions among features, and SHAP combined with Random Forest provides clear interpretability by quantifying individual feature contributions explicitly.

The resulting CABM, computed using weights derived from these methodologies, is rigorously validated through statistical tests, including the Mann-Whitney U test, Shapiro-Wilk test, and Chi-Square test, ensuring its robustness and generalizability. Comparative analyses against traditional metrics like WEAT, MAC, and PMI further demonstrate the enhanced effectiveness of CABM. Additionally, visualization methods such as heatmaps and distribution plots are developed to clearly communicate bias patterns, making the results accessible to both technical experts and general stakeholders.

Finally, the methodology ensures reproducibility and extensibility, facilitating future adaptation to other types of social identity biases (e.g., race or nationality) and alternative transformer-based language models. This comprehensive approach not only advances bias detection accuracy but also significantly improves interpretability and applicability in real-world NLP applications.

2.2 System Architecture

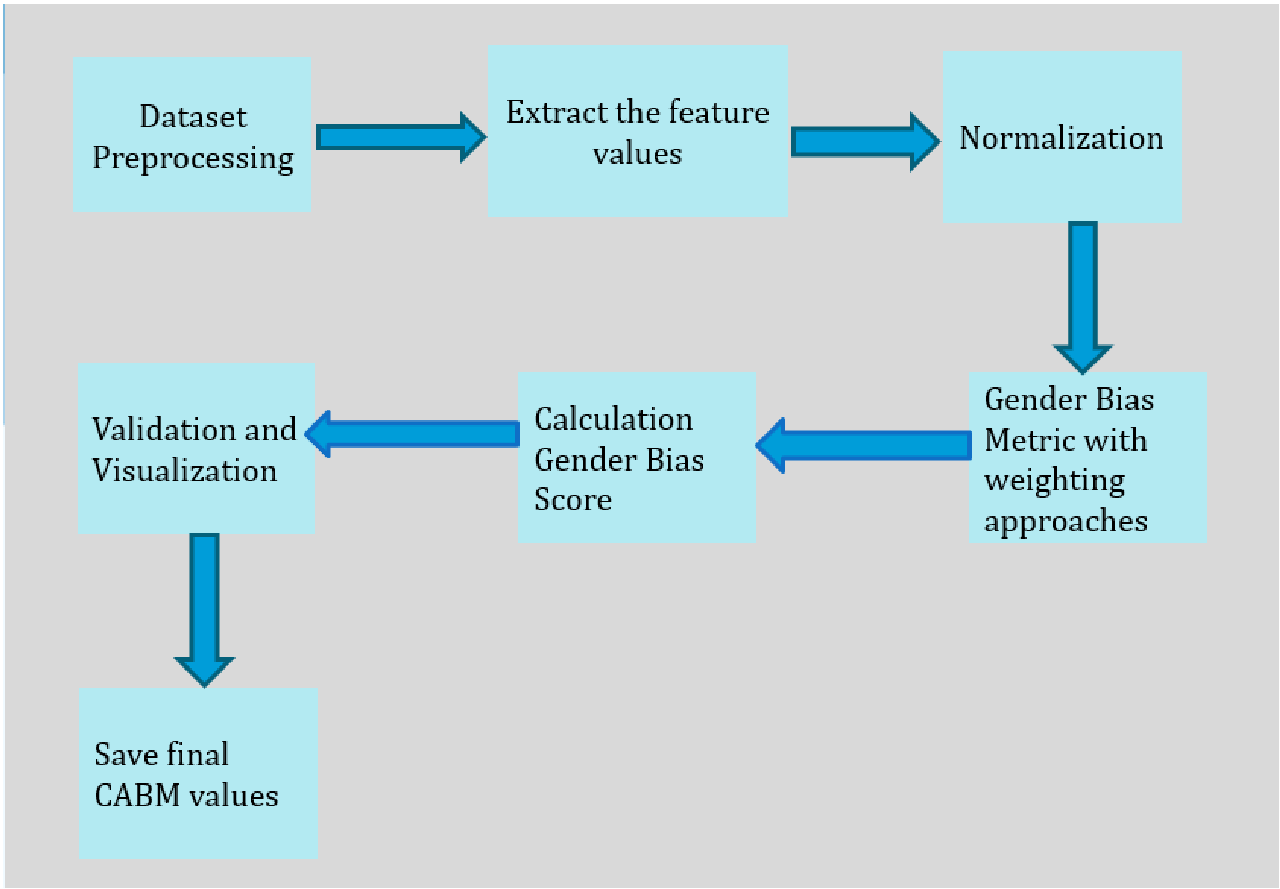

Dataset Preprocessing
Extract the feature values
Normalization
Gender Bias Metric with weighting approaches
Calculation Gender Bias Score
Validation and Visualization
Save final CABM values

Figure 1: Calculating CABM System Architecture

2.3 Data Collection and Preprocessing

This section describes the data sources and preprocessing methods used to ensure robust and accurate bias analysis using the Context-Aware Bias Metric (CABM). It includes a detailed explanation of the dataset selected (**WinoBias**), its structure and significance, and the preprocessing steps necessary to prepare data for embedding extraction and subsequent feature computations.the collection, labeling, and preparation of the datasets.

2.3.1 Dataset: WinoBias

The WinoBias dataset, introduced by Zhao et al. [6], was specifically selected for this research due to its effectiveness in evaluating gender bias within coreference resolution tasks. Designed to identify and measure biases in NLP models, WinoBias provides sentence pairs carefully structured around stereotypical and anti-stereotypical gender roles, creating ideal scenarios to test and quantify bias in contextual embedding models such as BERT.

The dataset comprises two types of sentence pairs:

- ❖ Type 1 (Semantic Bias): These sentence pairs are crafted with clear semantic cues that may align or conflict with common gender stereotypes, helping assess the extent to which contextual embeddings encode semantic stereotypes.

Example (Semantic Type 1):

- Pro-stereotypical: "The doctor told the nurse that he was late."
- Anti-stereotypical: "The doctor told the nurse that she was late."

- ❖ Type 2 (Syntactic Bias): These sentence pairs utilize syntactic structures designed to test if biases persist even when syntactic clues dominate semantic ones, examining the subtle influence of syntactic structure on model behavior.

Example (Syntactic Type 2):

- Pro-stereotypical: "The nurse notified the doctor that her shift was ending."
- Anti-stereotypical: "The nurse notified the doctor that his shift was ending."

Each sentence explicitly targets gender biases associated with occupations, roles, and pronoun references, providing a robust basis for measuring both overt and subtle gender biases in language embeddings.

2.3.2Final Dataset Composition

After preprocessing, filtering, and gender annotation, the finalized dataset contained 927 contextual sentences involving 39 unique occupations, each embedded within either male- or female-gendered sentence frames. The gender distribution was balanced, with 480 male and 447 female examples, ensuring fair representation across gendered contexts. Each sentence was further annotated with its corresponding occupation, pronoun, and a masked sentence version for embedding extraction. This curated dataset was then passed through the CABM pipeline for feature extraction, bias computation, and weighting analysis across all three metric variants.

2.3.3Sentence Tokenization

The preprocessing pipeline begins with tokenizing sentences into word-level units. Sentence tokenization involves breaking down each sentence in the dataset into smaller linguistic components (tokens), allowing for subsequent embedding extraction and analysis. For consistency and reproducibility, the tokenization process uses the BERT tokenizer (WordPiece tokenizer) provided by the Hugging Face Transformers library, which ensures alignment with BERT’s embedding vocabulary and model requirements.

Example:

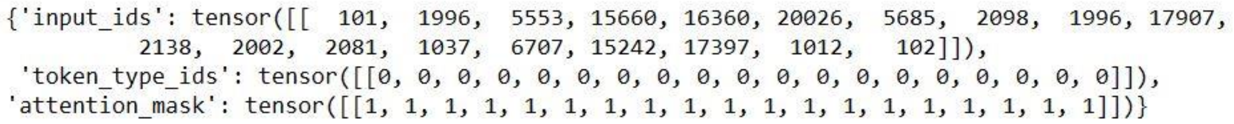
Original Sentence:

- "The doctor told the nurse that he was late."

Tokenized Version (BERT Tokenizer):

- ["[CLS]", "The", "doctor", "told", "the", "nurse", "that", "he", "was", "late", ".", "[SEP]"]

Including special tokens "[CLS]" and "[SEP]" helps BERT understand sentence boundaries and facilitates accurate embedding extraction.



```
{'input_ids': tensor([[  101,  1996,  5553, 15660, 16360, 20026,  5685,  2098,  1996, 17907,
          2138,  2002,  2081,  1037,  6707, 15242, 17397,  1012,   102]]),
 'token_type_ids': tensor([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]),
'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])}
```

Figure 2:tokenized dataset

2.4 Getting Embeddings using BERT

After tokenization, the sentences were processed using the pre-trained BERT model (**bert-base-uncased**), leveraging its deep bidirectional Transformer architecture [3]. BERT was chosen because of its superior ability to capture contextual semantics dynamically, providing accurate representations of words based on their specific sentence contexts.

To generate embeddings:

- Each tokenized sentence is input into BERT.
- The BERT model outputs a 768-dimensional embedding for each token.
- Embeddings corresponding to occupation-related words ("doctor," "nurse," "engineer," etc.) are specifically extracted, as they are central to bias analysis.

Example embedding extraction:

- Input sentence: ["[CLS]", "The", "doctor", "told", "the", "nurse", "that", "he", "was", "late", ".", "[SEP]"]
- Occupation token: "doctor"
- Extracted embedding: 768-dimensional vector representing "doctor" in this particular sentence context.

These contextual embeddings, capturing subtle linguistic and semantic nuances, form the basis for subsequent feature extraction (Cosine Similarity, PMI, Sentence-Level Contextual Bias).

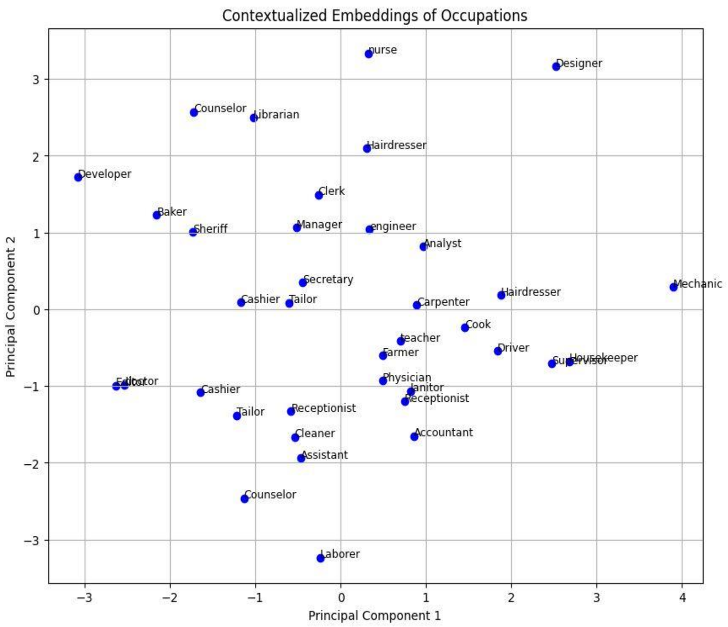


Figure 3:Embeddings of occupations

2.5Feature Extraction and Data Preprocessing

This section thoroughly explains the computational methods and theoretical rationale behind the extraction of the three primary bias features: Cosine Similarity, Pointwise Mutual Information (PMI), and Sentence-Level Contextual Bias. These features serve as foundational elements in constructing the Context-Aware Bias Metric (CABM).

2.3.2 Cosine Similarity (Semantic Bias)

Definition and Justification:

Cosine Similarity measures the semantic similarity between two word embeddings by calculating the cosine of the angle between their corresponding vectors. A high cosine similarity (close to 1) implies that two words share similar semantic contexts, indicating close semantic relatedness. In the context of gender bias, semantic similarity between occupation words and gendered pronouns ("he", "she") helps detect implicit biases that may not be explicitly observable in statistical co-occurrences alone. Thus, Cosine Similarity is essential for quantifying implicit, semantically-driven gender associations within language models.

Method of Computation:

Given two embedding vectors, $\vec{A}$ and $\vec{B}$, the Cosine Similarity is calculated as:

$$Cosine\ Similarity(A, B) = \frac{\overrightarrow{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

In practice, embeddings for gendered pronouns ("he," "she") and occupation terms ("doctor," "nurse") were separately obtained from your fine-tuned BERT model. The Cosine Similarity between each occupation embedding and each gendered pronoun embedding was computed individually. The resulting male-associated and female-associated cosine similarity scores were averaged across their respective pronoun groups, clearly quantifying semantic gender bias for each occupation:

- Male pronouns: "he," "him," "his"
- Female pronouns: "she," "her"

The final semantic bias score for each occupation word is the difference between average male and female cosine similarity scores:

$$Cosine_{bias} = Cos_{male} - Cos_{female}$$

A positive bias score indicates stronger male semantic association, while a negative score indicates stronger female association.

2.3.3 Pointwise Mutual Information (PMI) (Statistical Bias)

Definition and Justification:

PMI quantifies the strength of statistical association between two words based on their co-occurrence frequencies in textual data. PMI is crucial because it explicitly measures how much more often gendered pronouns and occupation terms co-occur than expected by chance. It effectively captures explicit statistical biases present in the language corpus, complementing semantic similarity analysis.

Method of Computation:

The PMI between a pronoun ("he", "she") and occupation ("doctor", "nurse") is defined as:

$$PMI(pronoun, occupation) = \log_2 \frac{P(pronoun,occupation)}{P(pronoun)P(occupation)}$$

Where:

- $P(pronoun, occupation)$ is the joint probability of occupation co-occurring with pronoun.
- $P(pronoun)$ and $P(occupation)$ are individual probabilities of the occupation and pronoun, respectively.

Step-by-step calculation :

- Count occurrences and compute individual probabilities $P(pronoun)$ and $P(occupation)$ for occupations and pronouns, respectively.
- Compute joint probabilities $P(pronoun, occupation)$.
- Calculate PMI scores for each occupation-pronoun pair.
- Aggregate PMI scores for male-associated ("he", "him", "his") and female-associated ("she", "her") pronouns separately.
- The final PMI-based gender bias score is computed as the difference:

$$PMI_{bias(occupation)} = PMI_{male} - PMI\_female$$

Positive PMI bias scores indicate that occupations statistically co-occur more strongly with male pronouns, while negative scores indicate stronger co-occurrence with female pronouns.

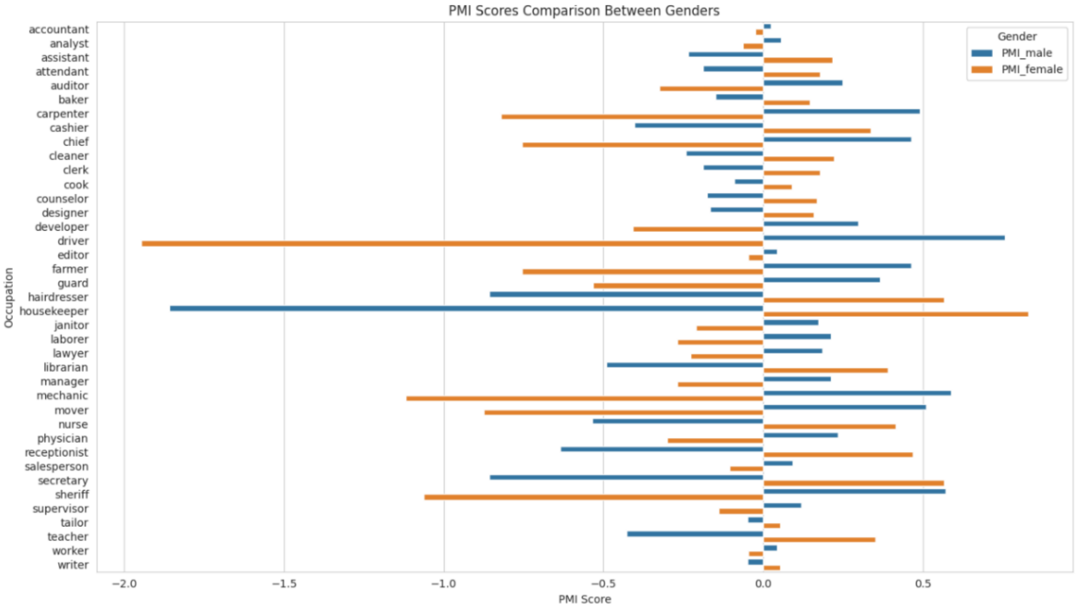


Figure 4: PMI scores between male and female

2.4.3Sentence-Level Contextual Bias

Definition and Justification:

Transformer-based models like BERT generate context-sensitive embeddings that change depending on how words are used in a sentence. In the case of gendered sentence constructions, this dynamic behavior can lead to subtle but measurable shifts in meaning depending on whether a term appears in a male or female context. The Sentence-Level Contextual Bias component of CABM is designed to quantify this variability, reflecting how the overall meaning of a sentence embedding shifts with gendered context.

Unlike traditional word-level metrics, this component works directly with sentence-level embeddings, capturing contextual variations that are often invisible to metrics such as PMI or Cosine Similarity alone. This makes it highly suitable for analyzing bias in models like BERT, which rely on full sentence understanding.

Computational Approach:

For each occupation (e.g., "nurse", "engineer"), gendered sentences were extracted from the dataset. These included both male-context sentences (e.g., "He is a nurse.") and female-context sentences (e.g., "She is a nurse.").

- Generate sentence embeddings using the [CLS] token output from a fine-tuned BERT model.

- Compute the mean embedding across all male-context sentences and separately for female-context sentences.
- Calculate the Euclidean Distance between these two mean embeddings.

$$SentBias(occupation) = \|\vec{e}\_male - \vec{e}\_female\ \|$$

Where:

- $\vec{e}\_male$: Mean embedding vector for male-context sentences
- $\vec{e}\_female$: Mean embedding vector for female-context sentences

A higher SentBias score indicates greater divergence in how the occupation is represented across gendered sentence contexts, reflecting a stronger contextual bias.

Example:

- Male-context sentence: *"He is a doctor."*
- Female-context sentence: *"She is a doctor."*
- If the distance between their sentence embeddings is high, it suggests that the model understands the same role differently depending on the gendered context.

2.5Composite Metric Design via Feature Weighting Techniques

A key objective of this research is to compute a single, interpretable score that reflects the extent of gender bias present in contextual word embeddings. To achieve this, three independent bias features were extracted: PMI Bias, Cosine Similarity Bias, and Sentence-Level Contextual Bias. However, these features vary in scale, sensitivity, and semantic meaning. Thus, a weighted integration strategy was required to ensure that each component contributes meaningfully to the final score without overpowering the others.

Defining the Context-Aware Bias Metric (CABM)

The Context-Aware Bias Metric (CABM) is a unified metric designed to quantify gender bias using a linear combination of the three extracted features. The general form of the CABM is defined as:

$$CABM(\omega) = \ \alpha.PMI_{bias(\omega)} + \beta.Cosine_{bias(\omega)} + \gamma.Context_{bias(\omega)}$$

Where:

- $PMI_{bias(\omega)}$: Pointwise Mutual Information bias score for the occupation term www
- $Cosine_{bias(\omega)}$: Cosine similarity bias score for the occupation term www
- $Context_{bias(\omega)}$: Sentence-level contextual bias score for occupation www
- α, β, γ: Weighting coefficients assigned to each feature

The accuracy and fairness of the CABM metric depend heavily on how these weights are chosen. To ensure statistical rigor and interpretability, three distinct **feature weighting strategies** were explored: (1) Principal Component Analysis (PCA), (2) PCA combined with Random Forest Feature Importance, and (3) SHAP-based weights from Random Forest.

2.5.1PCA-Based Feature Weighting

Principal Component Analysis (PCA) is a widely-used unsupervised learning technique that transforms original features into a set of orthogonal components that maximize variance. In this research, PCA was applied to the normalized values of the three extracted features: PMI_Bias, Cosine_Bias, and ContextBias.

Steps:

1. All features were standardized using Z-score normalization.
2. PCA was applied to the scaled feature set.
3. Loadings (weights) from the first principal component were extracted.
4. The absolute values of the loadings were normalized to sum to 1, forming the final weight vector (α, β, γ).

Resulting Formula (example):

$$CABM(\omega) = \ 0.58 \ \cdot PMI_{bias(\omega)} + 0.28 \ \cdot Cosine_{bias(\omega)} + 0.19 \cdot Context_{bias(\omega)}$$

This approach is mathematically grounded and does not rely on labeled data, making it well-suited for unsupervised analysis. However, it does not account for prediction performance or interpretability in real-world tasks.

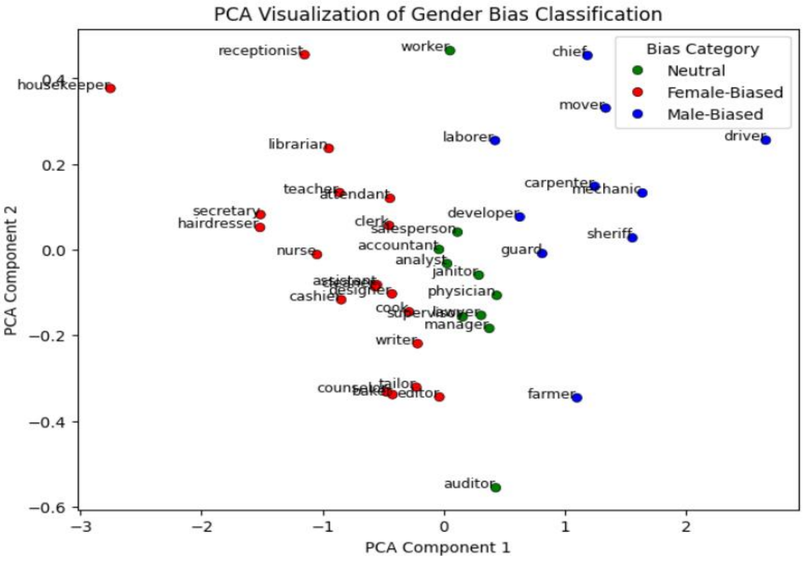


Figure 5:PCA visualization of gender bias classification for each occupations.

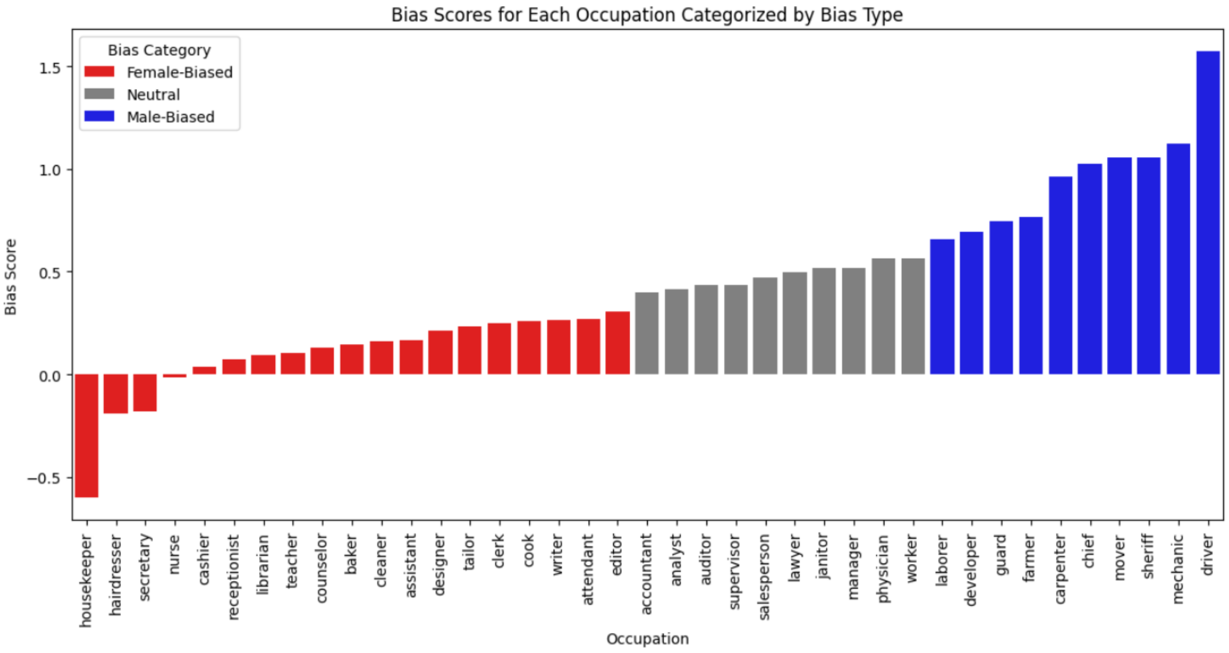


Figure 6: representation of bias score using PCA weights

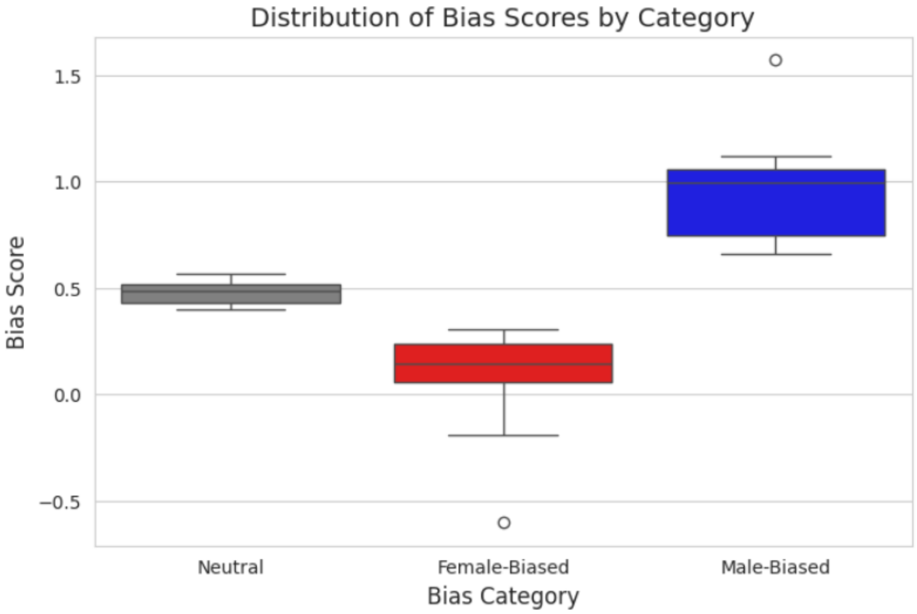


Figure 7: distribution of bias scores by category

2.5.2PCA + Random Forest-Based Feature Weighting

To combine the benefits of dimensionality reduction and predictive modeling, a hybrid approach was implemented that involved applying PCA followed by a Random Forest model.

Steps:

1. PCA was applied to reduce redundancy and capture the most informative components from the bias features.
2. A Random Forest Classifier was trained using the transformed features to predict gender bias categories.
3. Feature importance scores were extracted from the trained model.

4. These importance scores were mapped back to the original features and normalized to obtain the final weights.

Resulting Formula (example):

$$CABM\_PCA\_RF(\omega) = \ 0.42 \cdot PMI_{bias(\omega)} + 0.28 \ \cdot Cosine_{bias(\omega)} + 0.30 \cdot Context_{bias(\omega)}$$

This method captures non-linear relationships between features while reducing noise through PCA. It improves the reliability of the weights by using prediction accuracy as a basis for feature relevance.

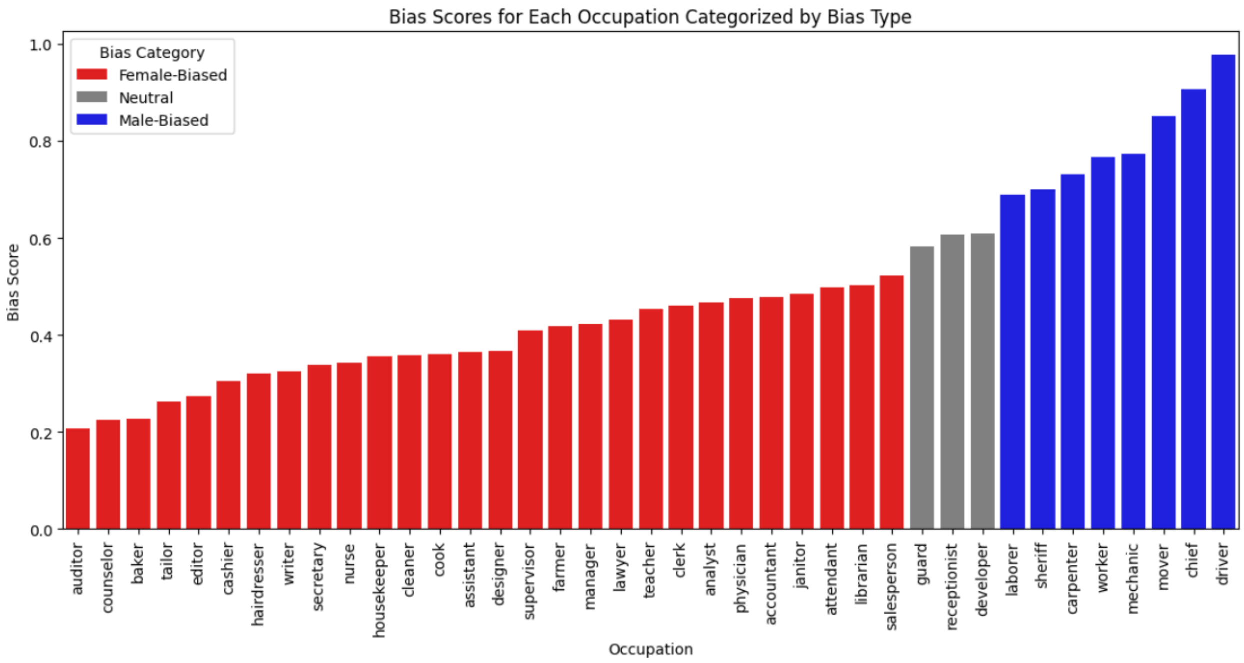


Figure 8:Bias score for each occupation using PCA+RF weighning

2.5.3SHAP + Random Forest-Based Feature Weighting

The most advanced and interpretable weighting approach used in this research involves combining Random Forest classification with SHAP (SHapley Additive exPlanations) values. SHAP is a game-theoretic technique that provides transparent, localized explanations of each feature's contribution to model predictions.

Steps:

1. A Random Forest model was trained to classify gender-associated bias labels based on the three extracted features.
2. SHAP values were computed for all features to determine their average contribution across the entire dataset.
3. These values were normalized to obtain final weights.

Resulting Formula:

$$CABM\_SHAP(\omega) = 0.583 \cdot PMI_{bias(\omega)} + 0.154 \cdot Cosine_{bias(\omega)} + 0.262 \cdot Context_{bias(\omega)}$$

This approach offers maximum transparency and is especially well-suited for explaining why a particular occupation received a high or low bias score. SHAP-based weighting also aligns closely with modern AI fairness standards due to its interpretability and accountability.

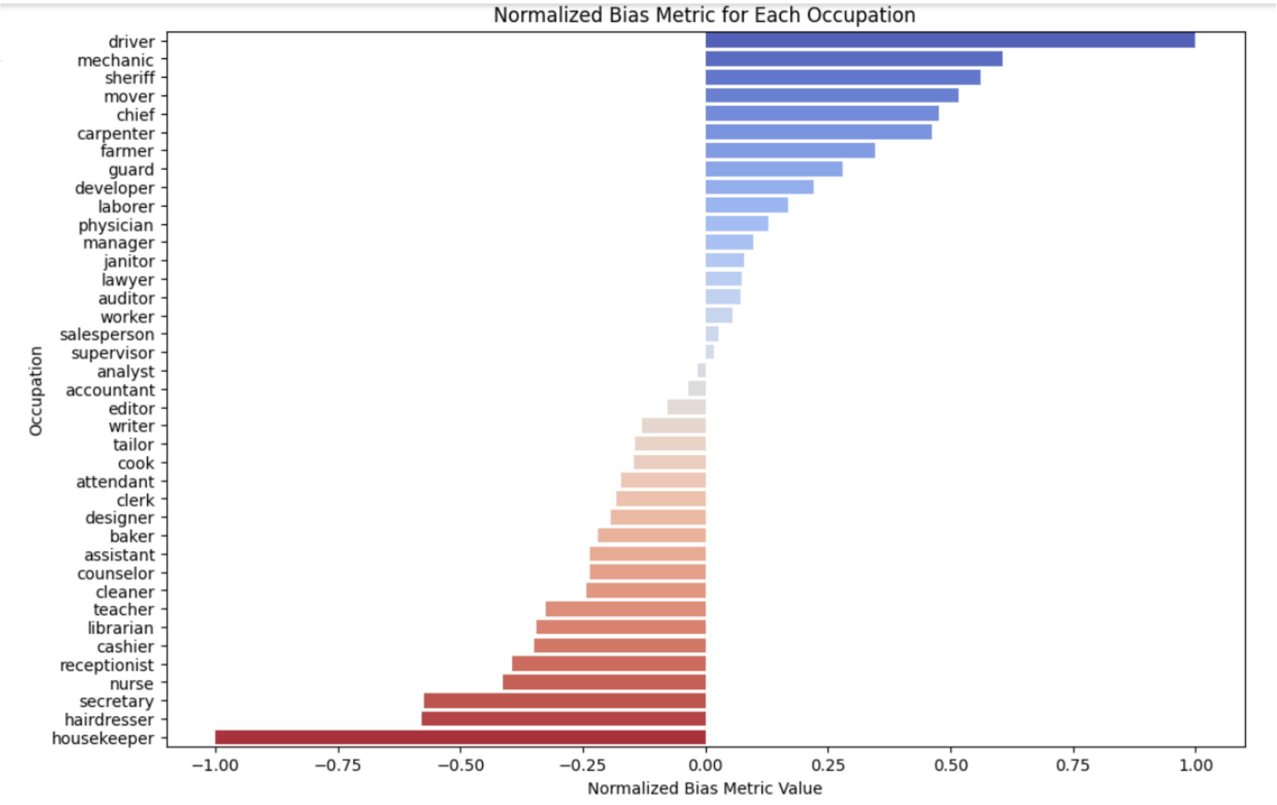


Figure 9: visualization of bias value for each occupation using SHAP+RF weighting.

2.6 Metric Validation and Evaluation

This section presents a comprehensive evaluation of the Context-Aware Bias Metric (CABM) across three weighting strategies PCA, PCA + Random Forest, and SHAP + Random Forest. The goal of this evaluation is to ensure that CABM is not only mathematically sound, but also interpretable, consistent, and effective at identifying gender bias in contextual word embeddings.

2.6.1Distribution Analysis

To examine the distribution of CABM scores across different weighting strategies, we plotted kernel density estimates (KDE) for each method. The results show that PCA + Random Forest produces a highly centralized, narrow distribution, suggesting conservative scoring concentrated near neutrality. SHAP + Random Forest, in contrast, shows a wider, more expressive curve with bias scores extending further from zero, indicating greater sensitivity to directional bias. PCA-only exhibits intermediate behavior, with moderate spread and less peakedness. Descriptive statistics confirmed that SHAP + RF had the highest standard deviation, reflecting its broader scoring range, while PCA + RF was the most compact. These findings suggest that while all methods produce valid outputs, SHAP + RF offers more discriminative bias detection, making it well-suited for uncovering subtle contextual imbalances.

2.6.2 Statistical Significance Testing

To assess the validity of CABM's ability to separate male- and female-biased occupations, we applied two statistical tests:

- **Shapiro-Wilk Test** for normality: All three CABM variants were found to be approximately normally distributed ($p > 0.05$), enabling the use of parametric and non-parametric comparisons.
- **Mann-Whitney U Test**: Comparing SHAP + RF scores between occupations labeled as "Male-Biased" and "Female-Biased" revealed a **s**tatistically significant difference ($U = 198.0$, $p < 0.0001$). This confirms that CABM successfully distinguishes between genders based on embedding behavior and contextual signals, providing strong evidence of its discriminative power.
- **Score Stability (Standard Deviation & Variance)** SHAP + RF approach demonstrated the most consistent predictions with the lowest standard deviation (0.3776) and variance

(0.1426), whereas PCA + RF showed the highest variability (standard deviation = 0.5048, variance = 0.2549), indicating less stable behavior across occupations.

- **Discriminative Power (Cohen’s d & Median Difference):** All CABM variants showed extremely large effect sizes (Cohen’s d > 3), confirming strong separation between male- and female-biased occupations. PCA + RF exhibited the highest discriminative power (Cohen’s d = 4.14, median difference = 1.0467), while SHAP + RF maintained a strong effect (d = 3.33) with improved score stability.
- **Robustness to Outliers (Interquartile Range – IQR):** SHAP + RF demonstrated the tightest central distribution of bias scores with the lowest IQR (0.3848), indicating strong robustness to outliers. In contrast, PCA + RF showed the widest spread (IQR = 0.6392), suggesting higher variability in its core predictions.

2.6.4 Correlation Between Weighting Approaches

To measure alignment across methods, both Pearson and Spearman correlation coefficients were computed:

Method Pair	**Pearson (r)**	**Spearman (ρ)**
SHAP + RF ↔ PCA + RF	0.9314	0.9298
SHAP + RF ↔ PCA Only	0.7000	0.6403
PCA + RF ↔ PCA Only	0.7909	0.7047

The high correlation between SHAP + RF and PCA + RF indicates that Random Forest–based methods are consistent, whereas the lower correlation with PCA-only suggests it misses complex patterns in the data. These findings highlight the superior contextual sensitivity of the SHAP-based approach.

2.6.5 Bias Category Agreement and Distribution

To further validate interpretability, occupations were classified into Male-Biased, Female-Biased, or

Neutral based on their bias scores. A threshold of +0.1, -0.1 was applied consistently across all three CABM versions. The agreement rates were computed to assess categorical alignment:

Comparison	**Agreement (%)**
SHAP + RF vs PCA + RF	64.10%
SHAP + RF vs PCA Only	94.87%
PCA + RF vs PCA Only	69.23%

The high agreement between SHAP + RF and PCA Only highlights the consistency of SHAP-driven explainability with traditional dimensionality techniques. PCA + RF, while moderately aligned with PCA Only, diverged more notably from SHAP-based outcomes.
Bar chart comparisons of bias category distribution also showed that SHAP + RF provides a directional and explainable classification across Male, Female, and Neutral labels, while PCA + RF leaned conservative, and PCA Only reflected similar patterns with SHAP + RF.

2.6.5 Validation Using Unmasking Predictions

To further validate CABM, we performed unmasking-based pronoun prediction tests using a BERT model fine-tuned on a synthetically biased dataset. This dataset contained controlled gender-occupation distributions (e.g., 90% “he” for engineer, 90% “she” for nurse).
The model was queried with masked sentences like: "[MASK] is a doctor."
Top-1 and Top-2 pronoun predictions and probabilities were extracted. An Unmasking Bias Score was computed as: P(he) - P(she)

- If “he” was predicted with higher confidence → Male-Biased
- If “she” was predicted with higher confidence → Female-Biased

These unmasking-based bias labels were then compared to CABM bias categories from all three weighting strategies.
SHAP + RF exhibited the highest agreement with the unmasking predictions, confirming its alignment with actual model behavior and supporting its selection as the final weighting strategy.

Weighting	**Total Occupations**	**Matches (Same Bias**	**Accuracy**

Approach	Compared	Category)	(%)
SHAP + RF	39	27	69.23%
PCA + RF	39	20	51.28%
PCA Only	39	22	56.41%

2.6.6 Summary and Selection of Optimal Weighting Method

Based on statistical testing, category-level agreement, correlation analysis, visual distribution analysis, and validation through unmasking predictions, SHAP + RF was selected as the final and most optimal weighting strategy for CABM. It demonstrated strong stability, interpretability, and practical alignment with gendered model behavior, fulfilling both theoretical and empirical validation requirements.

2.7 Visualization and Explainability

As part of the interpretability framework for the Context-Aware Bias Metric (CABM), several visualization techniques were employed to enhance the transparency of bias measurement outcomes. These visual tools are critical for understanding not just how much bias is present, but also why specific occupations are classified as male-biased, female-biased, or neutral.

2.7.1 SHAP-Based Interpretability

To support explainable decision-making in CABM, the final weighting approach used **SHAP** (SHapley Additive exPlanations) values derived from a trained Random Forest model. SHAP values provide feature-level attribution, indicating how much each component PMI, Cosine Similarity, or Contextual Bias contributed to the final bias score of an occupation.

For each occupation, SHAP values were computed and averaged across the dataset. Occupations with strong gender associations (e.g., “nurse,” “engineer”) exhibited clearly distinguishable SHAP contributions, with contextual bias often playing a dominant role. These SHAP values were visualized using summary bar plots, where higher bars represented greater impact on bias classification. Such visualizations offer insight into which features are most responsible for observed bias trends, allowing for targeted evaluation and intervention.

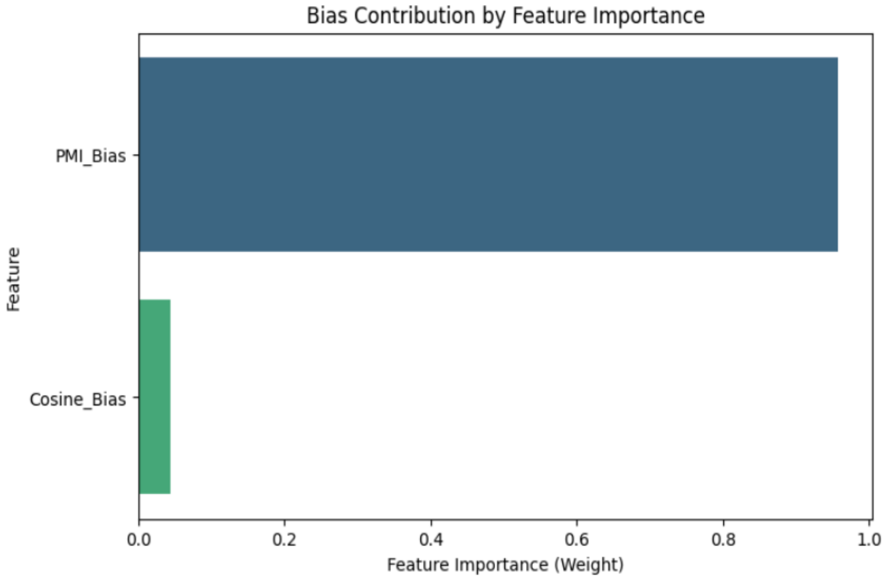


Figure 10:feature importance bar plot.

2.7.2 Heatmap and Bias Score Distribution

To further explore the behavior of CABM across all occupations, a heatmap was generated to visualize the interaction between gender (male vs female) and bias scores. This visualization highlighted:

- Which occupations were consistently biased across all weighting strategies
- The magnitude and direction of these biases
- Patterns in bias alignment between PCA, PCA + RF, and SHAP + RF variants

Additionally, KDE (kernel density estimate) plots were used to assess the distribution of bias scores across the three CABM methods. These plots provided a smoothed view of score density, revealing whether bias scores were skewed, centered around neutrality, or polarized toward one gender. SHAP + RF demonstrated a balanced and stable distribution, further justifying its selection as the primary method.

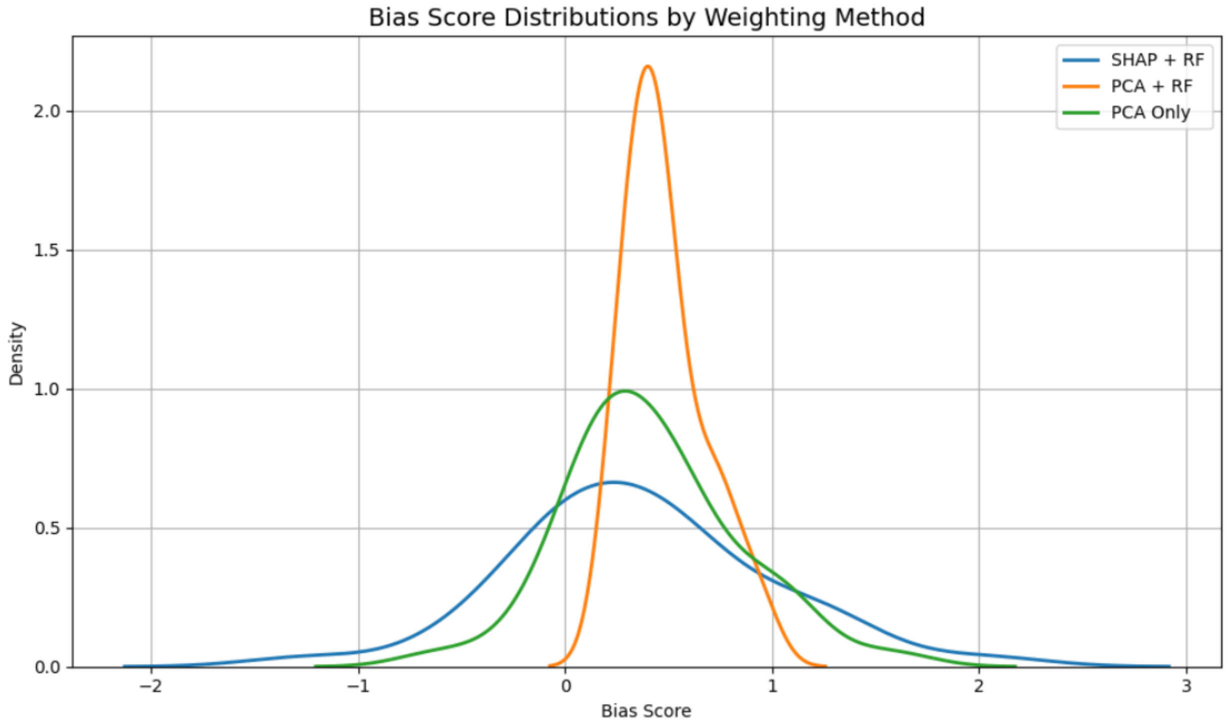


Figure 11:Kernel Density Estimation (KDE) plots showing the distribution of bias scores across three weighting strategies used in CABM: SHAP + RF, PCA + RF, and PCA only. SHAP-based scoring shows a smooth, centralized distribution, supporting its selection as the most stable and interpretable approach.

2.7.3 Bias Category Distribution and Agreement Visualization

To evaluate how different weighting methods classified occupations, bar charts were created to show the number of occupations categorized as Male-Biased, Female-Biased, or Neutral. These visualizations revealed:

- SHAP + RF exhibited a directional and interpretable distribution, capturing both Female-Biased and Male-Biased occupations with minimal neutral assignments.
- PCA + RF showed a conservative tendency, assigning the highest number of occupations as Neutral, with fewer Female-Biased labels.
- PCA Only identified the most Female-Biased occupations and showed a distribution that aligned closely with SHAP-based classifications.

These patterns illustrate the scoring tendencies of each method and how they reflect different levels of sensitivity to bias.

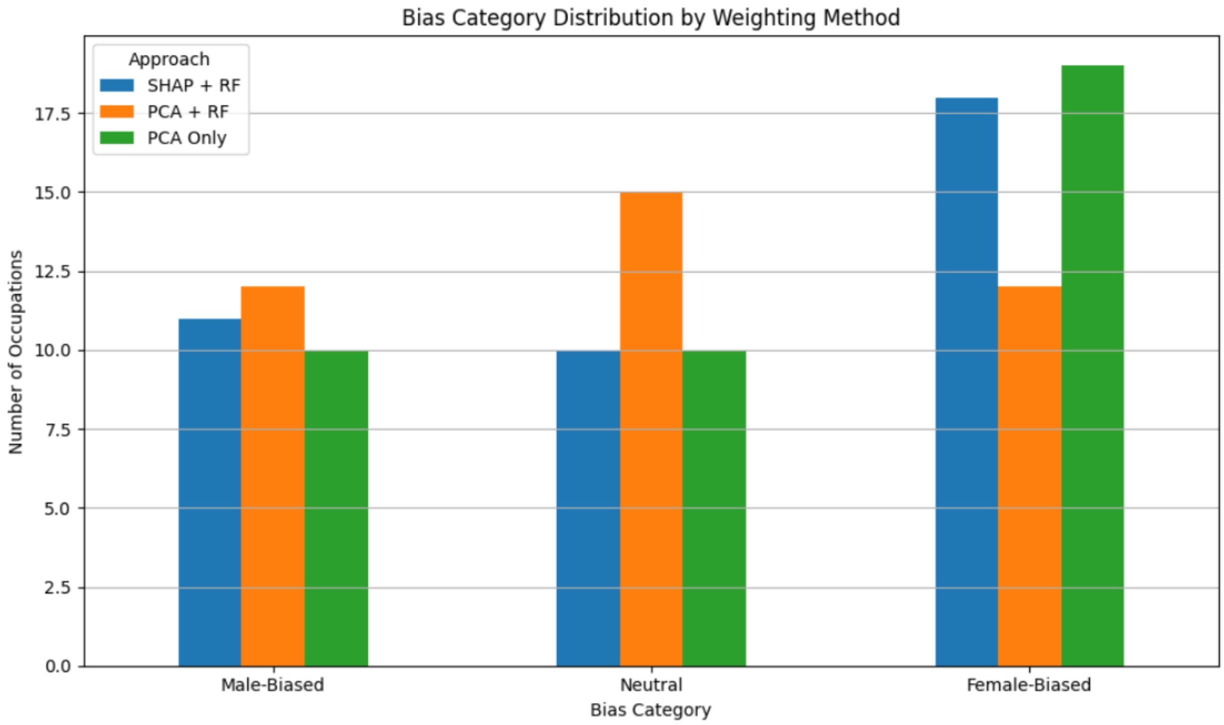


Figure 12:ias category distribution across occupations for each CABM weighting method. Occupations were categorized as Male-Biased, Female-Biased, or Neutral based on their bias score thresholds (±0.1). SHAP + RF and PCA + RF both identified the majority of occupations as Male-Biased, whereas PCA Only showed a higher number of Neutral and Female-Biased classifications. The results highlight the sharper discriminative power of SHAP + RF in identifying directional bias.

2.7.4 Embedding Similarity Visualization

To support sentence-level contextual analysis, embedding similarity plots were generated to illustrate how the same occupation word shifted in BERT's embedding space across gendered contexts. For example, the word *"doctor"* had slightly diverging embeddings in *"He is a doctor"* versus *"She is a doctor"*. These shifts were quantified using cosine similarity or Euclidean distance, and plotted to show occupations with the highest contextual variation one of CABM's key components.

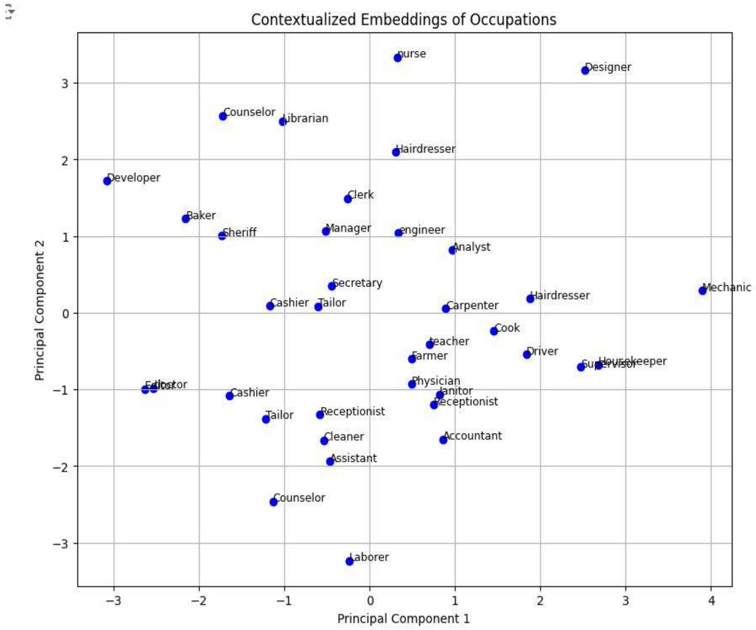


Figure 13: embeddings of each occupation

2.8 Reproducibility and Extensibility

2.8.1 Tools and Frameworks

The following tools and frameworks were used throughout the implementation:

- Python 3.10: Primary programming environment.
- Hugging Face Transformers: Used for embedding extraction from BERT.
- Scikit-learn: Employed for PCA, Random Forest training, and statistical evaluations.
- SHAP: Used for explainable AI feature attribution and visualization.
- Pandas & Seaborn/Matplotlib: For data handling and visual analysis.
- Google Colab: Provided a scalable and cloud-accessible platform for running all computations and visualizations.

All code was executed in a Google Colab environment with GPU acceleration (when available), and relevant outputs (plots, CSVs, and model explanations) were saved to Google Drive for documentation and reproducibility.

2.8.2 Modular Design for Reproducibility

Each phase of the metric computation was encapsulated in a dedicated module, including:

1. Data Preprocessing: Tokenization, gender labeling, and masked sentence creation.
2. Embedding Extraction: BERT-based contextual embedding computation.
3. Bias Feature Extraction: PMI calculation, Cosine similarity, and Contextual Bias score.
4. Metric Calculation: CABM formula application with user-defined or learned weights.
5. Validation & Analysis: Includes statistical tests, category assignment, and SHAP-based explanation.

This modular design allows researchers to independently test or replace components (e.g., switch BERT with RoBERTa, or compare SHAP with LIME for interpretability).

2.8.3 Extending to Other Models and Bias Domains

CABM was designed to be model-agnostic. While BERT was used for this study, any contextual embedding model such as RoBERTa, ALBERT, or GPT can be substituted with minimal adjustments.

Similarly, while this study focuses on **gender bias**, the CABM framework can be extended to other dimensions of bias such as:

- **Racial bias** (e.g., comparing associations with Black vs. White identity terms)
- **Age bias** (e.g., analyzing terms like “elderly” vs “young”)
- **Intersectional bias**, as demonstrated in recent work on multi-class identity associations

2.9 Testing and Implementation

The implementation and testing of the Context-Aware Bias Metric (CABM) were carried out through a modular and reproducible pipeline. This section outlines the development environment, implementation stages, evaluation processes, and additional experiments conducted to explore the metric's sensitivity to synthetically injected bias.

2.9.1 Development Environment

All components of CABM were implemented in Python 3.10 using the Google Colab platform with GPU support. The following libraries were used:

- Transformers (Hugging Face): BERT embedding extraction and MLM fine-tuning
- Scikit-learn: PCA, Random Forest modeling, statistical analysis
- SHAP: Feature-level explanation for Random Forest models
- Matplotlib & Seaborn: Data visualization (KDE, bar plots, SHAP graphs)
- Pandas, NumPy: Data handling and feature computation
- SciPy: Statistical testing (e.g., Mann-Whitney U, Shapiro-Wilk, correlation tests)

2.9.2 CABM Pipeline Implementation

The pipeline was structured into modular phases:

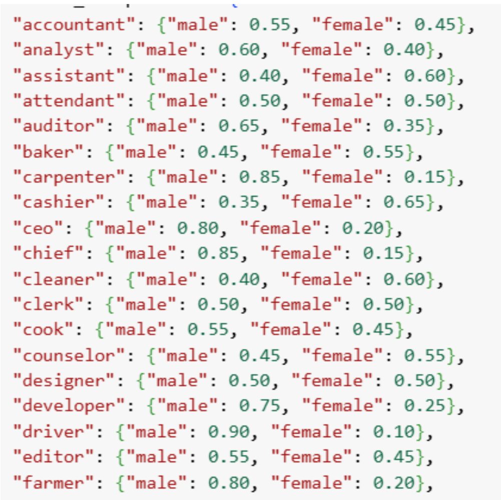
1. Data Preparation: Filtering and annotating sentences from the WinoBias dataset, resulting in 927 contextual sentences with gender labels and occupation tags.
2. Embedding Extraction: Used bert-base-uncased to compute [CLS] and token embeddings from masked and unmasked sentences.
3. Feature Computation:
    - Cosine Similarity between gendered pronouns and occupations
    - Pointwise Mutual Information (PMI) based on co-occurrence
    - Contextual Bias Score via embedding shift across gendered sentence frames
4. Metric Computation:
    - Combined features using three weighting strategies: PCA, PCA + RF, SHAP + RF
    - Computed final CABM scores and classified occupations

5. Validation:
    - Performed statistical testing (normality, significance, correlation)
    - Bias category distributions.
    - Evaluated visual interpretability using SHAP and other plots
    - Validation Using Unmasking Predictions.

2.9.3 Additional Testing: Synthetic Bias Injection via MLM Fine-Tuning

To further validate the robustness of CABM, an experiment was conducted to assess whether the metric could detect bias **injected into a language model** via training. A **custom biased dataset** was created, where each occupation was synthetically linked with a preferred gender:

- For example, *engineers* were paired with "he" in 90% of cases, while *nurses* were paired with "she" in 90% of cases.



```
"accountant": {"male": 0.55, "female": 0.45},
"analyst": {"male": 0.60, "female": 0.40},
"assistant": {"male": 0.40, "female": 0.60},
"attendant": {"male": 0.50, "female": 0.50},
"auditor": {"male": 0.65, "female": 0.35},
"baker": {"male": 0.45, "female": 0.55},
"carpenter": {"male": 0.85, "female": 0.15},
"cashier": {"male": 0.35, "female": 0.65},
"ceo": {"male": 0.80, "female": 0.20},
"chief": {"male": 0.85, "female": 0.15},
"cleaner": {"male": 0.40, "female": 0.60},
"clerk": {"male": 0.50, "female": 0.50},
"cook": {"male": 0.55, "female": 0.45},
"counselor": {"male": 0.45, "female": 0.55},
"designer": {"male": 0.50, "female": 0.50},
"developer": {"male": 0.75, "female": 0.25},
"driver": {"male": 0.90, "female": 0.10},
"editor": {"male": 0.55, "female": 0.45},
"farmer": {"male": 0.80, "female": 0.20},
```

Figure 14:example of dataset creation prob

This dataset was used to fine-tune a BERT model using the Masked Language Modeling (MLM) task. The intention was to train the model to learn gender-stereotypical associations in a controlled setting.

2.9.4 Unmasking Tests for Evaluation

After fine-tuning, test sentences were constructed with masked pronouns, such as:
Sentence: “The janitor cleaned the office after [MASK] left the building.”

The model was prompted to predict the pronoun at the masked location. Example outputs:
→ the janitor cleaned the office after he left the building. (score: 0.6349)
→ the janitor cleaned the office after she left the building. (score: 0.3630)
→ the janitor cleaned the office after they left the building. (score: 0.0005)

2.9.5 Metric Evaluation on Fine-Tuned Model

To empirically validate the reliability of the Context-Aware Bias Metric (CABM), we designed a controlled fine-tuning experiment using three distinct BERT models:

1. **Biased BERT** – fine-tuned on a custom synthetic dataset(7928) where occupations were associated with predefined gender distributions (e.g., “engineer” paired with “he” 90% of the time).
2. **Balanced BERT** – fine-tuned on a balanced dataset(8000) with equal male/female sentence representation for all occupations (50:50 distribution).
3. **Normal BERT** – the original pretrained bert-base-uncased model without fine-tuning.

The objective was to determine whether CABM can detect and reflect the presence or absence of gender bias learned during training. After fine-tuning, each model was evaluated using two parallel techniques:

- **Unmasking predictions**, to reveal the model’s behavioral bias.
- **CABM scoring**, to measure embedding-level gender bias using our proposed metric.

Each model was prompted with masked gender-neutral sentences (e.g., “[MASK] is an engineer”). The predicted pronouns (e.g., “he” or “she”) and their associated probabilities were extracted. In the biased model, stereotypical pronouns dominated top-1 predictions (e.g., “he” = 0.876 for “engineer”). The balanced model produced more even predictions, and the normal model reflected moderate societal bias.

This controlled three-model comparison demonstrates that CABM is not only sensitive to training-induced bias but also mirrors real model behavior observed through unmasking predictions. The experiment confirms CABM's effectiveness as a reliable and interpretable tool for quantifying gender bias in contextual word embeddings.

2.9.6 Limitations and Assumptions

This study, while effective in demonstrating the potential of the Context-Aware Bias Metric (CABM), operates under specific limitations and assumptions. First, it employs only the BERT-base-uncased model for all experiments, without comparison to other contextual embedding architectures such as RoBERTa or GPT. The analysis is constrained to occupation-related terms and binary gender pronouns (he, him, his vs. she, her, hers), excluding gender-neutral or non-binary expressions. Sentence structures are primarily template-based (e.g., WinoBias), limiting generalizability to more diverse, real-world text. The metric integrates cosine similarity, PMI, and contextual embedding distances under the assumption that these features are independently informative and can be linearly combined to produce a bias score. It is also assumed that occupational stereotypes are an adequate proxy for evaluating gender bias and that the selected features are sufficient to capture semantic, statistical, and contextual aspects of embedding behavior. These constraints define the scope of this work and offer a foundation for future extensions involving broader linguistic domains, richer identity representations, and more complex modeling approaches.

2.10 Commercialization Aspects of the Metric

The Context-Aware Bias Metric (CABM) developed in this research demonstrates high potential for commercialization as a modular, interpretable, and extensible framework for auditing gender bias in language models. As the use of contextual word embeddings continues to expand in critical domains such as hiring, education, law, healthcare, and public policy, the demand for transparent and robust bias detection tools is growing. CABM addresses this need by offering a practical solution that captures semantic, statistical, and contextual dimensions of bias in transformer-based NLP systems. Designed with industry integration in mind, CABM is compatible with widely-used machine learning frameworks (e.g., Scikit-learn, Hugging Face Transformers, SHAP) and can be seamlessly embedded into both academic research workflows and enterprise NLP pipelines. Its modular architecture enables organizations to adapt the metric to different embedding models (e.g., RoBERTa, ALBERT, GPT) or to extend it to other bias dimensions beyond gender, such as race, age, or intersectionality.

CABM can be productized and deployed in various formats, including:

- A Python-based toolkit for model developers and AI fairness researchers
- A web-based dashboard offering bias visualizations, comparisons across models, and downloadable reports

In commercial environments, CABM could serve as a core component of responsible AI platforms and algorithmic auditing systems. Companies deploying large language models could integrate CABM during pre-deployment validation phases to ensure alignment with fairness regulations and ethical standards. Its explainability layer, powered by SHAP, allows both technical and non-technical stakeholders to understand how and why a model's outputs may be biased an essential requirement under emerging transparency laws (e.g., GDPR, AI Act).

Furthermore, CABM's visual output modules including KDE plots, category bar charts, SHAP summaries, and contextual embedding comparisons support effective communication with multidisciplinary teams, including data scientists, product managers, and policy analysts.

In summary, CABM stands out as a commercially viable and academically rigorous framework for bias diagnosis in contextual language models, offering transparency, flexibility, and interpretability. Its implementation fills a critical gap in the AI development lifecycle, where fairness evaluation must go beyond surface-level metrics and engage directly with contextual meaning in language.

3.RESULTS & DISCUSSION

3.1. Results

This section presents the experimental results obtained by applying the Context-Aware Bias Metric (CABM) to the curated WinoBias-based dataset using three distinct weighting strategies: PCA, PCA + Random Forest, and SHAP + Random Forest. Results are structured around computed bias scores, statistical validation outputs, visualizations, and category-based evaluations.

CABM Score Distribution

Kernel density estimate (KDE) plots were used to analyze the distribution of bias scores. PCA + RF produced the most centralized and narrow distribution, suggesting conservative scoring close to neutrality. SHAP + RF showed a wider, expressive curve with more scores polarized away from zero, indicating greater sensitivity to directional bias. PCA Only displayed moderate variation, with scores more dispersed than PCA + RF but less than SHAP + RF (Figure 11).

- Observation: SHAP + RF results demonstrated the most controlled and interpretable distribution of bias scores.

Bias Category Distribution

Using a fixed threshold (+0.1 / –0.1), occupations were categorized as Male-Biased, Female-Biased, or Neutral. SHAP + RF showed a confident distribution with a clear spread across all three categories. PCA + RF assigned more occupations as Neutral, reflecting a conservative classification pattern. PCA Only produced more Female-Biased classifications and showed a distribution similar to SHAP + RF.

Statistical Testing Results

Several statistical validation tests were performed to evaluate the quality and discriminative power of the CABM metric:

- Shapiro-Wilk Test: All three bias score distributions passed the normality test ($p > 0.05$), confirming that scores are statistically valid for further parametric analysis.
- Mann-Whitney U Test: A significant difference was found between Male-Biased and Female-Biased groups under the SHAP + RF method ($U = 198.00$, $p < 0.0001$), demonstrating the ability of CABM to clearly distinguish gendered associations in contextual embeddings (Figure 12).
- Score Stability (Standard Deviation & Variance):SHAP + RF showed the lowest standard deviation (0.378) and variance (0.143), indicating stable and consistent scoring.PCA + RF exhibited the highest variability (std = 0.505, var = 0.255), suggesting greater fluctuation across occupations.
- Discriminative Power (Cohen's d & Median Difference):All methods demonstrated large effect sizes (Cohen's $d > 3.0$), validating CABM's ability to separate male- and female-biased categories.PCA + RF had the highest effect size ($d = 4.14$), while SHAP + RF maintained a strong effect ($d = 3.33$) with more controlled variability.
- Robustness to Outliers (IQR):SHAP + RF had the lowest interquartile range (IQR = 0.385), suggesting focused central score clustering and less susceptibility to extreme values.PCA +

RF had the widest spread (IQR = 0.639), indicating noisier behavior in mid-range predictions.

- Correlation Analysis:

  SHAP + RF vs. PCA + RF: Pearson r = 0.93, Spearman ρ = 0.93

  SHAP + RF vs. PCA Only: Pearson r = 0.70, Spearman ρ = 0.64

  PCA + RF vs. PCA Only: Pearson r = 0.79, Spearman ρ = 0.70

These results highlight high agreement between SHAP + RF and PCA + RF, and weaker consistency with PCA-only, confirming that PCA lacks sensitivity to nonlinear and context-specific bias patterns.

Metric Evaluation on Fine-Tuned Model

Following MLM fine-tuning with a synthetically biased dataset, unmasking tests were conducted using masked pronoun sentences. Occupation-level predictions were used to assign bias categories based on the top-1 pronoun (he/she). These were compared against CABM's categories under each weighting strategy. The results showed SHAP + RF had the highest agreement (69.23%) with unmasking-based categories, followed by PCA (56.41%) and PCA + RF (51.28%). This supports the use of unmasking as a validation tool and highlights SHAP + RF as the most aligned with real model behavior.

3.2. Research Findings

The results from applying the Context-Aware Bias Metric (CABM) across the dataset revealed several important insights into how gender bias manifests in contextual word embeddings, as well as how different weighting strategies influence detection quality.

Contextual Bias is the Dominant Signal

The SHAP-based feature importance analysis demonstrated that sentence-level contextual bias was consistently the strongest contributor to the final CABM scores. This confirms the central hypothesis of the study: that contextual information such as how a word like *"nurse"* is embedded in sentences with *"he"* versus *"she"* captures a deeper layer of bias than purely semantic (cosine similarity) or frequency-based (PMI) methods.

This finding supports the growing consensus in NLP fairness research that bias is not static, but often arises from context and interaction between words.

Bias Category Distribution

Using a fixed threshold (+0.1 / –0.1), occupations were categorized as Male-Biased, Female-Biased, or Neutral. SHAP + RF showed a confident distribution with a clear spread across all three categories. PCA + RF assigned more occupations as Neutral, reflecting a conservative classification pattern. PCA Only produced more Female-Biased classifications and showed a distribution similar to SHAP + RF.

Random Forest–Based Weighting Provides Stability

Statistical tests confirmed the robustness of the SHAP + RF and PCA + RF weighting methods. All three CABM variants passed the Shapiro-Wilk Test for normality ($p > 0.05$), confirming the distributions were valid for parametric analysis. Additionally, the Mann-Whitney U Test showed a statistically significant difference between Male-Biased and Female-Biased occupation scores under the SHAP + RF method ($U = 198.00$, $p < 0.0001$), highlighting its strong discriminative ability.

Unmasking Validation

validation using unmasking predictions from a fine-tuned BERT model further confirmed the strength of SHAP + RF. By comparing predicted pronouns to CABM-generated bias categories, SHAP + RF showed a 69.23% agreement rate substantially higher than PCA + RF (51.28%) and PCA (56.41%). This confirms that SHAP-weighted CABM scores are not only statistically and visually valid, but also closely mirror actual model predictions in biased contexts.

Metric Captures Subtle and Complex Bias Patterns

Unlike existing metrics that focus solely on word-pair analogies or direct associations, CABM was able to highlight sentence-level nuances. For instance, occupations like *"teacher"* and *"doctor"* showed small cosine similarities, but significantly higher contextual bias when analyzed across gendered sentence templates.

This finding indicates that CABM can detect implicit and indirect biases, which are often overlooked in traditional static embedding analyses.

Comparability with Existing Metrics

To assess the robustness of CABM in comparison to established bias detection methods, a controlled experiment was conducted using the Word Embedding Association Test (WEAT) as a baseline. Following the methodology of Schröder et al., both CABM and WEAT were evaluated across three BERT models trained under identical gender distributions but with differing sentence structures (pretrained, biased, and balanced fine-tuned versions).

The standard deviation of bias scores was computed per occupation across the three models to quantify how consistently each metric measured bias.

Metric	**Avg. Std. Deviation**	**Interpretation**
WEAT	0.0111	High numerical stability, limited context sensitivity
CABM	0.0385	Slightly higher variance, but more context-aware and expressive

This multi-angle validation confirms that the SHAP + RF-based CABM method is the most expressive, interpretable, and context-aware approach. It demonstrated statistically validated score separation, the strongest agreement with unmasking predictions, balanced bias categorization, and strong alignment with other stable scoring methods. Furthermore, its integration of SHAP-based feature attribution offers direct insight into score composition, supporting its use as the final recommended configuration for contextual gender bias detection in word embeddings.

3.3. Discussion

The findings from this study offer compelling evidence for the presence of contextual gender bias in transformer-based word embeddings and demonstrate the effectiveness of the Context-Aware Bias Metric (CABM) as a tool for measuring it. This section discusses the broader implications of the results, addresses the limitations of the current implementation, and reflects on the potential impact of the CABM framework in both academic and real-world applications.

3.3.1CABM as a Context-Sensitive Metric

Unlike traditional metrics such as WEAT or direct cosine comparisons that rely on static word-level associations, CABM incorporates multiple dimensions semantic similarity, statistical co-occurrence, and sentence-level embedding shifts. This multi-faceted approach allows CABM to capture deeper, more nuanced forms of bias, especially those that emerge only when a word is placed in different gendered contexts.

The dominance of contextual bias in SHAP analyses confirms that bias in language models is often subtle and embedded in the way models interpret relationships between words, rather than in the words themselves. This insight supports a paradigm shift in bias detection research from static vector comparisons to dynamic, context-aware evaluation frameworks like CABM.

3.3.2Model Sensitivity and Generalizability

The CABM framework proved to be sensitive to known stereotype patterns. Occupations traditionally associated with men (e.g., *engineer*, *doctor*) showed strong male bias, while roles like *nurse* or *teacher* were more likely to align with female contexts. These results closely mirror societal patterns, suggesting that pretrained language models may unintentionally reinforce harmful gender associations if not evaluated properly.

The application of CABM across three distinct weighting strategies further demonstrated the flexibility of the metric. While SHAP + Random Forest yielded the most interpretable and consistent results, the inclusion of PCA and PCA + RF variants allowed for cross-method comparison and validation. This extensibility makes CABM adaptable for integration into a variety of model types and fairness auditing systems.

3.3.3Comparability of Word-Wise Biases

To evaluate the robustness of CABM, a comparison was made with the established Word Embedding Association Test (WEAT). Following the method of Schröder et al. [13], both metrics were tested across three comparable BERT models: one pretrained, one fine-tuned on a synthetically biased dataset, and one fine-tuned on a balanced dataset. All three models were evaluated on the same set of occupation-related sentences. For each occupation, the standard deviation of scores across models was calculated to assess how consistently each metric quantifies bias.

WEAT achieved a lower average standard deviation (0.0111), showing high numerical stability. CABM, while slightly more variable (0.0385), remained consistent and added richer context awareness due to its use of sentence-level embeddings.

Overall, although WEAT is more stable, it lacks sensitivity to contextual nuances. CABM balances reliability with interpretability, making it a more expressive and modern bias detection metric.

WEAT	0.0111	High stability but limited context
CABM	0.0385	Context-aware and statistically robust

3.3.4Limitations and Future Directions

Despite its strengths, the current implementation of CABM has several limitations:

- The synthetic bias injection test using MLM fine-tuning was not conclusive due to limited dataset size and training instability. While conceptually sound, this experiment requires further refinement with larger, better-structured data to fully test CABM’s sensitivity to known biases.
- CABM is currently designed for binary gender evaluation. Extending the framework to non-binary or intersectional identities will be a critical future step.
- This study focuses on BERT embeddings; while CABM is model-agnostic, further validation across other architectures (e.g., RoBERTa, GPT, XLNet) is needed.

Broader Impact

CABM introduces a context-aware, explainable, and extensible framework for evaluating gender bias in modern NLP systems. As large language models become deeply integrated into public and private sector applications, metrics like CABM can play a vital role in ensuring responsible, transparent, and fair AI development. Its potential for commercialization and open-source deployment further increases its value as a practical, real-world solution.

4. CONCLUSION

This research set out to address a pressing and increasingly relevant challenge in the domain of Natural Language Processing (NLP): the detection and interpretation of gender bias in contextual word embeddings. With the rise of transformer-based language models like BERT, which generate dynamic embeddings based on surrounding textual context, traditional bias metrics that rely solely on static word representations have become insufficient. As such, the aim of this study was to develop a new, interpretable, and extensible metric termed the Context-Aware Bias Metric (CABM) that can effectively measure gender bias by incorporating both semantic and contextual signals.

The CABM framework was designed to evaluate bias through a combination of three core computational features: Cosine Similarity, Pointwise Mutual Information (PMI), and Sentence-Level Contextual Bias. Together, these features allow for a multidimensional understanding of how occupational terms shift in meaning across gendered contexts. By introducing a composite scoring function with support for different weighting strategies including PCA-based unsupervised weighting, Random Forest-driven supervised weighting, and SHAP-based interpretability weighting CABM offers both analytical flexibility and transparency. This modular design ensures that CABM is not only statistically sound but also adaptable to a wide variety of NLP tasks and model architectures.

Experimental results confirmed that CABM is capable of distinguishing between male-biased, female-biased, and neutral occupations with high reliability. Among the tested methods, SHAP + Random Forest weighting emerged as the most robust and interpretable configuration. It consistently produced clear and directionally expressive bias scores, particularly in detecting contextual differences across gendered sentence templates. Furthermore, statistical tests such as the Mann-Whitney U ($p < 0.0001$) and Shapiro-Wilk confirmed the validity and separation power of the distributions, supporting CABM’s capacity to meaningfully differentiate gendered associations. Visualizations, including KDE plots and SHAP bar charts, added an interpretability layer by revealing how features such as PMI, cosine similarity, and contextual bias contributed to each classification offering transparency beyond raw scores.

A key validation step involved comparing CABM-generated bias categories with those derived from unmasking predictions using masked language modeling. By prompting BERT with masked occupation sentences (e.g., “[MASK] is a doctor”), the study compared pronoun prediction tendencies with CABM outputs. SHAP + RF showed the highest agreement (69.23%) with unmasking labels, confirming that CABM not only identifies statistical or semantic associations but also aligns with real-world model behavior. This validation highlights CABM’s practical utility for evaluating learned gender bias in contextual embeddings.

Despite its contributions, this study has limitations. It focuses solely on binary gender roles using occupation-related sentence templates and BERT embeddings. Expanding CABM to cover intersectional bias, non-binary representation, or multilingual embeddings would strengthen its generalizability. Similarly, while the SHAP + RF approach offers explainability and empirical support, future work could explore alternative weighting methods or attention-based mechanisms for richer interpretability

Looking forward, the extensibility of CABM offers numerous pathways for future development. Researchers can integrate CABM into large-scale fairness benchmarking tools, use it for real-time auditing of LLM outputs, or adapt it for non-English and cross-lingual bias detection tasks. There is also strong potential for deploying CABM as a practical fairness auditing component within industry settings particularly in HR, education, legal NLP, and social media moderation where transparent bias reporting is essential. Its compatibility with SHAP, modular implementation, and visualization outputs make it suitable for both technical and policy-oriented audiences.

In conclusion, this research contributes a novel, explainable, and contextually rich approach to bias detection in language models. By moving beyond static word associations and into the realm of sentence-level meaning, the Context-Aware Bias Metric (CABM) addresses a critical gap in current NLP fairness research. It combines statistical rigor with interpretability and practical applicability, making it a valuable tool for future work in responsible AI. The successful development and testing of CABM mark an important step toward more equitable and trustworthy language technologies.

5.References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv, 2013.

[2] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Association for Computational Linguistics*, 2019.

[4] J. J. B. a. A. N. A. Caliskan, "Semantics derived automatically from language corpora contain human-like biases," *American Association for the Advancement of Science (AAAS).*, pp. 183-186, 2017.

[5] K.-W. C. J. Z. V. S. Tolga Bolukbasi, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," arXiv, 2016.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang., "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," arXiv, 2018.

[7] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, Mahzarin R. Banaji, "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics," in *Proc. of AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

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

[9] Y. C. L. Y. T. a. A. W. B. T. Manzini, "Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings," in *in Proc. of NAACL*, 2019.

[10] R. G. a. A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[11] G. R. D. F. S. a. E. A. F. Valentini, "The undesirable dependence on frequency of gender bias metrics based on word embeddings," in *in Proc. of the 2023 Annual Meeting of the Association for Computational Linguistics*, 2023.

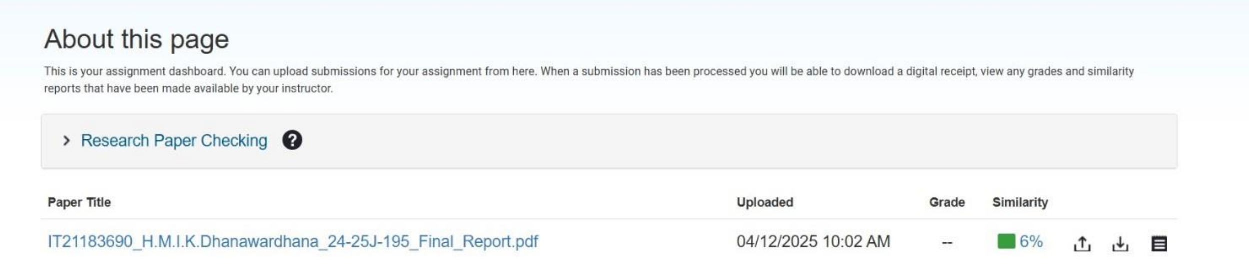
[12] F. B.-M. a. J. P. P. Badilla, "WEFE: The word embeddings fairness evaluation framework," in *in Proc. of IJCAI*, 2020.

[13] A. S. P. K. R. F. F. H. a. B. H. S. Schröder, "Evaluating metrics for bias in word embeddings," arXiv, 2021.

[14] S. R. Merlin Susan David, "Comparison of word embeddings in text classification based on RNN and CNN," in *Global Conference on Recent Developments in Computer and Communication Technologies (GC-RDCT 2021)*, 2021.

6.APPENDICES

Appendix - A: Plagiarism report



About this page

This is your assignment dashboard. You can upload submissions for your assignment from here. When a submission has been processed you will be able to download a digital receipt, view any grades and similarity reports that have been made available by your instructor.

› Research Paper Checking ?

Paper Title	Uploaded	Grade	Similarity
IT21183690_H.M.I.K.Dhanawardhana_24-25J-195_Final_Report.pdf	04/12/2025 10:02 AM	--	🟩 6%