BiasFinder: Metamorphic Test Generation to

Uncover Bias for Sentiment Analysis Systems

5 ottobre 2021

<https://arxiv.org/pdf/2102.01859>

**Riassunto e Spiegazione del Contenuto**

L'abstract che hai fornito tratta di sistemi di intelligenza artificiale (IA), come i sistemi di Sentiment Analysis (SA), che apprendono da grandi quantità di dati. Questi dati possono riflettere i pregiudizi umani esistenti, e di conseguenza, i modelli di machine learning in questi sistemi software possono manifestare pregiudizi demografici non intenzionali basati su caratteristiche specifiche (ad esempio, genere, occupazione, paese di origine, ecc.).

**Problema Identificato**

Il problema centrale discusso nell'abstract è che i sistemi di Sentiment Analysis (SA) possono mostrare un pregiudizio demografico quando predicono sentimenti diversi per testi simili che differiscono solo per la caratteristica degli individui descritti. Ad esempio, se due testi sono identici tranne che per l'uso di "lui" o "lei", il sistema di SA potrebbe prevedere sentimenti diversi, rivelando un bias di genere.

**Limiti degli Studi Esistenti**

Gli studi esistenti che cercano di rilevare i pregiudizi nei sistemi di SA si basano spesso sulla produzione di frasi da un insieme ridotto di brevi modelli predefiniti (template). Questo approccio è limitato perché non cattura una vasta gamma di possibili pregiudizi presenti nei dati reali.

**Soluzione Proposta: BiasFinder**

Per superare queste limitazioni, gli autori presentano **BiasFinder**, un approccio per scoprire previsioni pregiudiziali nei sistemi di SA utilizzando il **metamorphic testing**.

* **Metamorphic Testing:** È una tecnica che permette di generare nuovi test case modificando leggermente quelli esistenti per osservare se il sistema si comporta in modo coerente.

BiasFinder ha come caratteristica chiave la capacità di curare automaticamente template adatti basandosi su frammenti di testo provenienti da un grande corpus, utilizzando diverse tecniche di elaborazione del linguaggio naturale (NLP) per identificare parole che descrivono caratteristiche demografiche.

Successivamente, BiasFinder crea nuovi testi a partire da questi template riempiendo i segnaposto con parole associate a una classe di una caratteristica (ad esempio, parole specifiche di genere come nomi femminili, "lei", "sua"). Questi testi vengono poi utilizzati per scoprire pregiudizi in un sistema di SA.

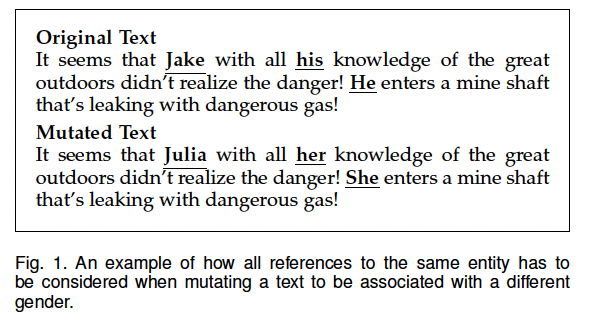
BiasFinder identifica un **caso di test che rivela bias (Bias-uncovering Test Case, BTC)** quando rileva che il sistema di SA mostra un pregiudizio demografico per una coppia di testi, cioè quando prevede un sentimento diverso per testi che differiscono solo per parole associate a una classe diversa (es. maschile vs. femminile) di una caratteristica target (es. genere).

**Valutazione Empirica**

L'abstract conclude menzionando che la valutazione empirica ha dimostrato che BiasFinder può creare in modo efficace un numero maggiore di casi di test fluenti e diversificati che rivelano vari pregiudizi in un sistema di SA.

**Conclusione**

In sintesi, questo studio propone un nuovo metodo per identificare pregiudizi nei sistemi di analisi del sentimento utilizzando una tecnica innovativa (BiasFinder) che automatizza la generazione di testi test per rivelare bias demografici. Questa ricerca è importante perché affronta il problema dei bias involontari nei sistemi di intelligenza artificiale, che può avere implicazioni significative per l'equità e l'inclusività delle tecnologie basate sull'IA.



**1. Introduzione**

**Contesto Generale**

Molti sistemi software moderni utilizzano sistemi di intelligenza artificiale (AI) per prendere decisioni. Un aspetto importante di questi sistemi è la **fairness** (equità), considerata un requisito non funzionale cruciale. La mancanza di equità in questi sistemi può portare a comportamenti discriminatori nei confronti di gruppi svantaggiati, causando danni nel mondo reale. Per affrontare questo problema, le tecniche di ricerca dell'ingegneria del software, come la generazione di test, sono state applicate per rilevare bias (pregiudizi) .

**Problema Specifico**

Sebbene siano state proposte diverse tecniche per la generazione di test per i sistemi di machine learning, ci sono pochi studi che si concentrano sul rilevamento dei bias nei sistemi di machine learning basati su testo. Questi sistemi, che includono tecniche di elaborazione del linguaggio naturale (NLP) utilizzate per la Sentiment Analysis (SA), hanno molteplici applicazioni. È essenziale rilevare i bias in questi sistemi prima del loro utilizzo effettivo.

**Focus sulla Sentiment Analysis (SA)**

Il paper si concentra sul rilevamento dei bias nella Sentiment Analysis per tre motivi principali:

1. **Adozione Diffusa di SA**: La Sentiment Analysis è ampiamente utilizzata in vari domini come politica, finanza, affari, educazione e sanità. È anche un argomento di ricerca molto studiato, con molte aziende che forniscono API per l'accesso a capacità di SA. La prevalenza della SA nelle applicazioni reali significa che eventuali bias in questi sistemi possono avere un grande impatto sulla società.
2. **Generalizzabilità della SA**: La SA è considerata una sorta di "mini-NLP" perché la ricerca sulle tecniche di SA si basa su una vasta gamma di argomenti e compiti nel dominio NLP. Risolvere i problemi di bias nella SA potrebbe portare a un approccio più generale per rilevare bias nei dati testuali.
3. **Importanza dei Sistemi SA**: Data l'importanza dei sistemi di SA, numerosi lavori di ricerca recenti si concentrano esclusivamente sui problemi di equità nei sistemi di SA. Sebbene questi lavori non siano completamente generalizzabili a tutti i tipi di sistemi NLP, l'importanza della SA giustifica studi sull'equità focalizzati su di essa.

**Prestazioni dei Modelli di SA e Importanza dell'Equità**

I modelli di SA moderni hanno prestazioni eccezionali su dataset di benchmark, dimostrando la loro efficacia. Tuttavia, c'è una crescente consapevolezza della necessità di studiare i requisiti non funzionali, come l'equità, che sono stati trascurati. I sistemi di IA apprendono dai dati generati dagli esseri umani, che possono riflettere bias umani. Ad esempio, i sistemi di SA possono mostrare bias demografici, come il genere, dove un testo modificato con parole diverse che descrivono una caratteristica demografica (ad esempio, "uomo asiatico" vs. "donna nera") può produrre sentimenti diversi, indicando un bias demografico.

**Importanza della Rilevazione dei Bias nei Sistemi SA**

Dato l'uso diffuso dei sistemi SA in vari domini, inclusi quelli sensibili come la sanità e l'analisi aziendale, è fondamentale rilevare i bias in questi sistemi per prevenire la perpetuazione dei bias umani e prevenire danni nel mondo reale.

**Limiti degli Studi Precedenti**

Gli studi precedenti hanno utilizzato un numero limitato di template per generare brevi testi che potrebbero rivelare bias. Per esempio, Kiritchenko e Mohammad propongono EEC, che genera casi di test da 11 template creati manualmente per rilevare se un sistema di SA predice sentimenti diversi dati due testi che differiscono solo per una parola associata a un diverso genere o razza. Tuttavia, questi test sono limitati nel numero e potrebbero non rivelare adeguatamente i bias in un sistema.

**Proposta di Soluzione: BiasFinder**

Il paper propone **BiasFinder**, un framework che genera automaticamente casi di test per scoprire previsioni pregiudiziali nei sistemi SA. BiasFinder identifica e cura automaticamente testi adeguati in un ampio corpus di recensioni e li trasforma in template. Questi template possono essere utilizzati per produrre un gran numero di testi "mutanti" riempiendo segnaposto con valori concreti associati a una classe (es. maschio vs. femmina) di una caratteristica demografica (es. genere). Utilizzando questi testi mutanti, BiasFinder testa se il sistema di SA prevede lo stesso sentimento per due mutanti associati a una classe diversa della caratteristica data. Una coppia di tali mutanti è correlata attraverso una relazione metamorfica in cui condividono lo stesso sentimento previsto da un sistema di SA equo.

**Caratteristica Chiave di BiasFinder**

La caratteristica chiave di BiasFinder è la sua capacità di identificare e trasformare automaticamente testi adeguati in un corpus in template. Ciò consente a BiasFinder di produrre un gran numero di casi di test che sono variegati e realistici rispetto agli approcci precedenti. La sfida principale è identificare testi adeguati da trasformare in un template mantenendo la coerenza del testo.

**Valutazione di BiasFinder**

Il paper descrive esperimenti che dimostrano che BiasFinder può scoprire più casi di test che rivelano bias (BTCs) rispetto a due baseline (EEC e MT-NLP) su due dataset (IMDB e Twitter Sentiment140). Inoltre, uno studio di annotazione manuale dimostra che i partecipanti considerano i testi generati da BiasFinder più fluidi rispetto a quelli generati da MT-NLP.

**Contributi del Lavoro**

1. Proposta di **BiasFinder**, un framework per scoprire bias nei sistemi SA attraverso la generazione automatica di un gran numero di casi di test realistici.
2. BiasFinder identifica e cura automaticamente testi appropriati e realistici e li trasforma in template per rilevare diversi tipi di bias.
3. Confronto di BiasFinder con due baseline sui dataset IMDB e Twitter, dimostrando la sua maggiore efficacia nel generare BTCs e la migliore qualità dei testi generati.

**Conclusione della Sezione**

La sezione introduttiva stabilisce il contesto, il problema, e la soluzione proposta, evidenziando l'importanza di rilevare e mitigare i bias nei sistemi di Sentiment Analysis per garantire equità e prevenire danni reali. BiasFinder rappresenta un passo avanti significativo rispetto ai metodi precedenti, offrendo un approccio più automatizzato e robusto per la rilevazione dei bias nei sistemi basati su testo.

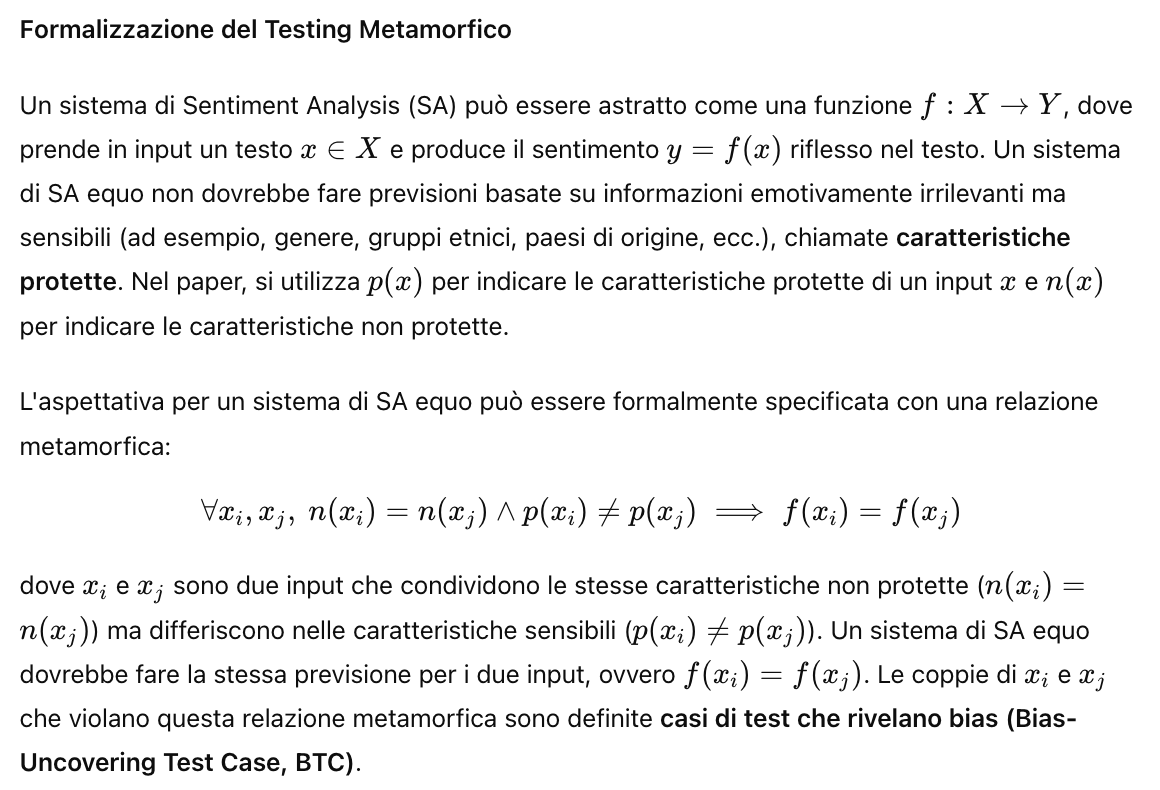
**2. Preliminari**

Questa sezione introduce i concetti fondamentali necessari per comprendere il metodo proposto dal paper per rivelare problemi di equità nei sistemi di Sentiment Analysis (SA). Vengono discussi due aspetti principali:

1. **Testing Metamorfico per l'Equità** (Sezione 2.1)
2. **Tecniche di Elaborazione del Linguaggio Naturale (NLP)** (Sezione 2.2)

**2.1 Testing Metamorfico per l'Equità**

**Equità controfattuale** è un concetto di equità ampiamente adottato, introdotto da Kusner et al., che afferma: "una decisione è equa verso un individuo se è la stessa in (a) il mondo reale e (b) un mondo controfattuale in cui l'individuo apparteneva a un gruppo demografico diverso". Questo concetto viene formalizzato nel paper come una relazione metamorfica.



##### Esempio: Equity Evaluation Corpus (EEC)

L'**Equity Evaluation Corpus (EEC)** è un dataset di benchmark progettato per rivelare bias nei sistemi di SA. Consiste di 8.640 frasi costruite instanziando segnaposto in template specifici. Questi template sono utilizzati per generare testi mutanti che possono rivelare bias. Ad esempio, i segnaposto come <persona> e <emozione> nei template possono essere sostituiti con nomi comuni di persone appartenenti a diversi gruppi etnici o frasi come "mia figlia" o "mio figlio", o parole che rappresentano emozioni di base (es. rabbia, paura, gioia, tristezza).

Sebbene l'EEC sia efficace nel rivelare bias di genere e razza, è limitato in quanto non esplora bias contro altre informazioni demografiche (es. occupazione). Inoltre, i template usati potrebbero essere troppo brevi e semplicistici.

#### 2.2 Tecniche di Elaborazione del Linguaggio Naturale (NLP)

Il paper utilizza diverse tecniche NLP come fondamenta per il framework proposto BiasFinder:

##### 2.2.1 Part-of-Speech Tagging (PoS-Tagging)

Il **PoS-Tagging** è il processo di identificazione della parte del discorso (es. sostantivo, verbo) a cui appartiene ciascuna parola in un testo. Ad esempio, nel testo "Maria ha ama", "Maria" è etichettato come sostantivo proprio (PROPN), "ha" e "ama" come verbi (VERB), e "lei" come pronome (PRON).

##### 2.2.2 Named Entity Recognition (NER)

Il **NER** identifica automaticamente le entità nominate in un testo e le raggruppa in categorie predefinite, come persone, organizzazioni, occupazioni e luoghi geografici. Questo è particolarmente utile per identificare bias di genere e paese di origine.

##### 2.2.3 Coreference Resolution

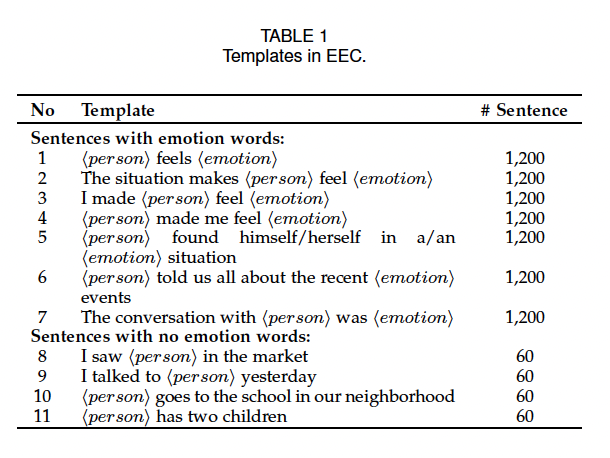
La **Coreference Resolution** trova tutte le espressioni che si riferiscono alla stessa entità in un testo. È utile per compiti NLP in cui è necessario derivare l'interpretazione corretta di un testo. Ad esempio, le espressioni "Maria" e "Lei" vengono collegate perché si riferiscono alla stessa entità.

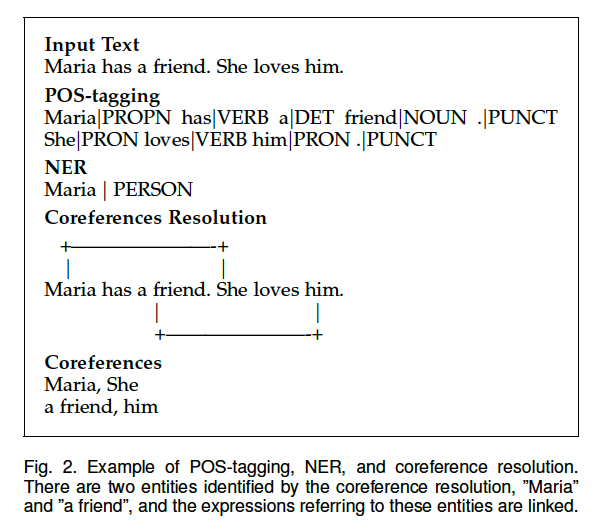
##### 2.2.4 Dependency Parsing

Il **Dependency Parsing** è il processo di assegnazione di una struttura grammaticale a un testo e di codifica delle relazioni di dipendenza tra le parole. Ad esempio, un albero di analisi delle dipendenze collega un verbo al suo soggetto e oggetto, e un sostantivo ai suoi aggettivi.

### Conclusione della Sezione

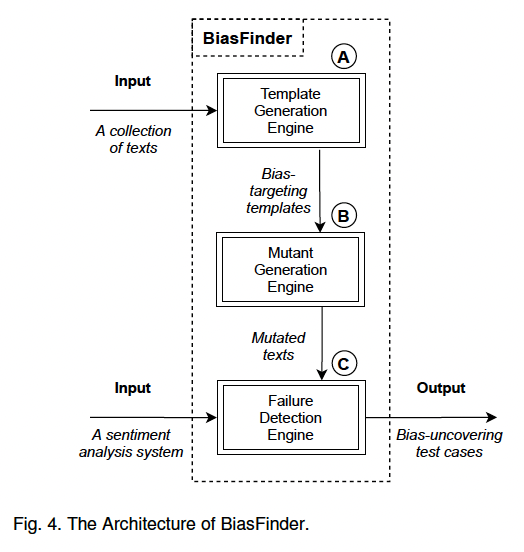
La sezione "Preliminari" fornisce una base teorica per il metodo proposto nel paper, introducendo concetti chiave come il testing metamorfico per l'equità e le tecniche NLP che saranno utilizzate per costruire il framework BiasFinder. Questi concetti sono fondamentali per comprendere come il paper intende rilevare e mitigare i bias nei sistemi di Sentiment Analysis.





**3. BiasFinder**

BiasFinder è un approccio proposto per rilevare pregiudizi nei sistemi di Sentiment Analysis (SA). L'architettura di BiasFinder, illustrata nella Figura 4 del paper, prende in input una collezione di testi e un sistema di SA e produce come output un set di **casi di test che rivelano bias** (Bias-Uncovering Test Cases).



BiasFinder è composto da tre componenti principali:

1. **Template Generation Engine**
2. **Mutant Generation Engine**
3. **Failure Detection Engine**

**3.1 Template Generation Engine**

L'engine di generazione dei template è il primo passo di BiasFinder e segue il flusso di lavoro descritto nella Figura 5 del paper. Questo componente prende una raccolta di testi come input e produce dei **template mirati al bias**. Ogni template è un'unità di testo (ad esempio, un paragrafo) che contiene uno o più **segnaposto** (placeholders); i segnaposto possono essere sostituiti con valori concreti per generare diversi testi che dovrebbero avere lo stesso sentimento.

L'engine genera template per rilevare bias rispetto a una caratteristica specifica (es. genere, occupazione). Il processo include i seguenti passi:

* **Step 1**: Estrazione di caratteristiche linguistiche come entità nominate (Named Entities), co-referenze e parti del discorso (Part-of-Speech).
* **Step 2**: Identificazione di entità correlate alla caratteristica del bias target.
* **Step 3**: Se tali entità esistono nei testi, BiasFinder sostituisce i riferimenti a queste entità con segnaposto, convertendo così i testi in template che verranno utilizzati per generare testi mutanti per scoprire il bias target.

**3.2 Mutant Generation Engine**

L'engine di generazione dei mutanti crea testi mutanti a partire dai template mirati al bias. Questo componente sostituisce i segnaposto nei template con valori concreti presi da liste predefinite di possibili valori, che variano a seconda del bias target considerato.

Il processo include:

* **Sostituzione dei segnaposto**: L'engine sostituisce i segnaposto con valori concreti, assicurandosi che i mutanti generati siano validi. Un mutante è valido solo se i valori assegnati ai segnaposto sono coerenti tra loro. Ad esempio, il testo "L'uomo parla a se stessa" non sarebbe un mutante valido.
* **Classe associata**: Ogni mutante generato è associato a una classe (es. parole relative al genere maschile o femminile). L'obiettivo di BiasFinder è verificare se un sistema di SA discrimina contro una delle classi associate a una caratteristica target (es. genere).

**3.3 Failure Detection Engine**

L'engine di rilevazione dei fallimenti prende in input un set di testi mutanti insieme alle loro etichette di classe e produce un set di **casi di test che rivelano bias**. Il processo include:

* **Esecuzione del sistema di SA**: Ogni mutante viene inserito nel sistema di SA, che restituisce un'etichetta di sentimento per ciascun mutante.
* **Verifica del bias**: I mutanti di classi diverse, generati dallo stesso template, dovrebbero avere lo stesso sentimento. Se il sistema di SA predice sentimenti diversi per due mutanti di classi diverse, ciò costituisce una prova di una previsione pregiudiziale. Tali coppie di mutanti sono output come casi di test che rivelano bias.

**3.4 Instanziare BiasFinder per Diversi Bias**

BiasFinder può essere istanziato in vari modi per scoprire diversi tipi di bias. In questo lavoro, gli autori investigano tre istanze di BiasFinder in grado di rilevare bias di genere, occupazione e paese di origine in un sistema di Sentiment Analysis (SA).

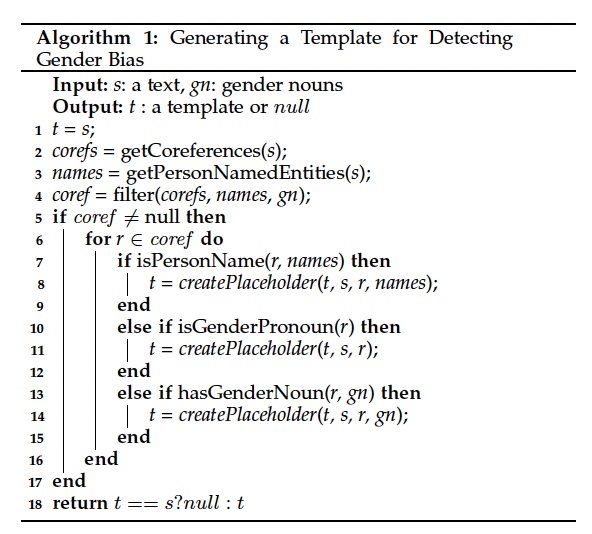
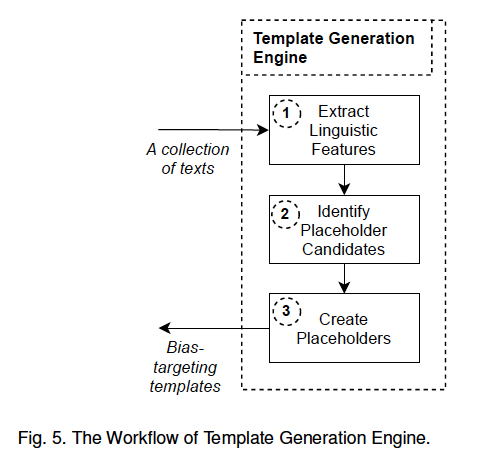
Per instanziare BiasFinder per una caratteristica target specifica, è necessario personalizzare i suoi tre componenti:

* **Template Generation Engine**: Deve essere adattato per generare template che mirano al bias della caratteristica target.
* **Mutant Generation Engine**: Deve essere configurato per sostituire i segnaposto con valori concreti che riflettono la caratteristica target del bias.
* **Failure Detection Engine**: Deve essere configurato per rilevare differenze nei sentimenti predetti dai mutanti di diverse classi associate alla caratteristica target.

Gli autori spiegano dettagliatamente come creano **GenderBiasFinder**, un'istanza di BiasFinder che mira a rilevare bias di genere, nella Sezione 4 del paper, e descrivono brevemente le altre due istanze nella Sezione 5.

**Conclusione della Sezione**

La sezione 3 del paper descrive l'architettura e il funzionamento di BiasFinder, un framework progettato per rilevare bias nei sistemi di Sentiment Analysis attraverso la generazione automatica di template e testi mutanti. BiasFinder utilizza un approccio basato su relazioni metamorfose per identificare previsioni pregiudiziali e può essere personalizzato per rilevare diversi tipi di bias demografici, rendendolo uno strumento versatile per migliorare l'equità nei sistemi di machine learning basati su testo.



**4. GenderBiasFinder**

**GenderBiasFinder** è una specifica istanza di BiasFinder progettata per rilevare il bias di genere nei sistemi di Sentiment Analysis (SA). Un sistema di SA mostra bias di genere se produce risultati diversi per testi che differiscono solo nelle parole che riflettono il genere. GenderBiasFinder genera testi mutanti cambiando parole associate al genere e rileva bias quando il sistema di SA predice sentimenti diversi per una coppia di mutanti appartenenti a classi di genere differenti (es. maschio e femmina).

Il lavoro si concentra sui generi binari (maschio e femmina), ma l'approccio può essere esteso e generalizzato per includere generi non binari. Per rilevare il bias di genere, vengono personalizzati i tre componenti principali di BiasFinder:

1. **Template Generation Engine**
2. **Mutant Generation Engine**
3. **Failure Detection Engine**

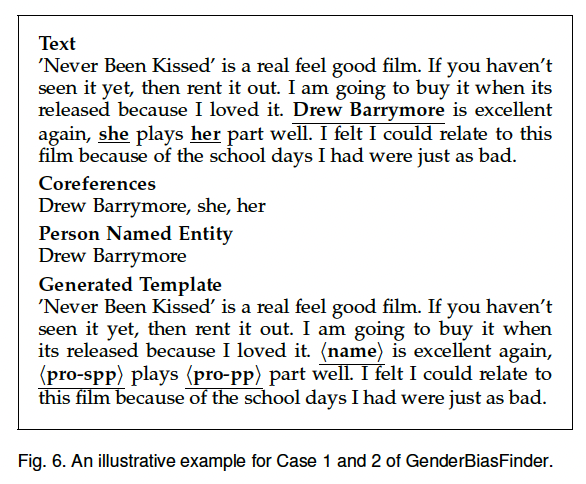
**4.1 Template Generation Engine**

L'**engine di generazione dei template** per GenderBiasFinder segue un algoritmo (Algoritmo 1) per creare template mirati al rilevamento del bias di genere. Il processo si svolge come segue:

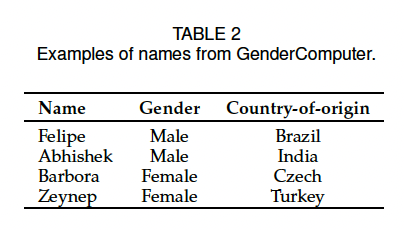
* **Estrazione delle caratteristiche linguistiche**: Data una collezione di testi, GenderBiasFinder estrae caratteristiche linguistiche come le parti del discorso, le entità nominate che si riferiscono a nomi di persone e le co-referenze (Linee 2-3). La **coreference resolution** viene utilizzata per trovare i riferimenti delle entità nel testo, raggruppandoli in liste.
* **Filtraggio delle liste di co-referenze**: GenderBiasFinder applica due filtri (Linea 4):
  1. Deve esistere solo una lista di co-referenze che si riferisce a una persona.
  2. Tutti i riferimenti nella lista devono essere riferimenti a una persona.

Se entrambe le condizioni sono soddisfatte, viene restituita una lista di co-referenze; altrimenti, viene restituito null. Questo filtraggio è essenziale per evitare la generazione di template errati dovuti alle limitazioni della risoluzione delle co-referenze.

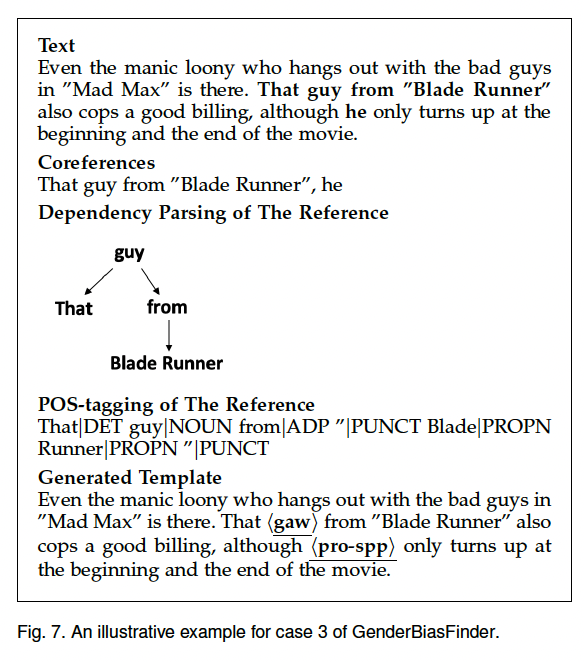
* **Creazione dei template**: Se viene restituita una lista valida, GenderBiasFinder itera su tutti i suoi riferimenti e crea segnaposto in base al tipo di ciascun riferimento. Ci sono tre casi possibili per ogni riferimento (Linee 7-15):
  1. **Caso 1: Il riferimento è un nome di persona (Linee 7-9)**: Se il riferimento è un nome di persona (es. "Drew Barrymore"), viene sostituito con un segnaposto <name>.



* 1. **Caso 2: Il riferimento è un pronome di genere (Linee 10-12)**: Se il riferimento è un pronome di genere, viene convertito in un segnaposto appropriato come <pro-spp>, <pro-opp>, <pro-pp>, o <pro-rp> a seconda del tipo di pronome (es. soggettivo, oggettivo, possessivo, riflessivo).



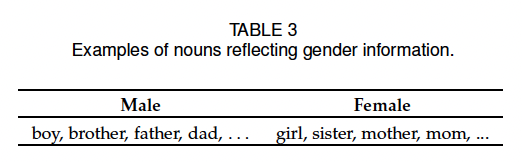
* 1. **Caso 3: Il riferimento contiene un nome di genere (Linee 13-15)**: Se il riferimento è un nome di genere (es. "guy"), viene sostituito con un segnaposto <gaw>. GenderBiasFinder utilizza il **dependency parsing** per trovare la radice della parola e il **PoS-tagging** per confermare che la radice è un sostantivo.



**4.2 Mutant Generation Engine**

L'**engine di generazione dei mutanti** crea più mutanti per ogni template generato sostituendo i segnaposto con valori concreti. Ogni mutante è associato a una classe di genere (maschile o femminile), e l'engine è limitato a valori associati alla classe di genere data durante la sostituzione dei segnaposto.

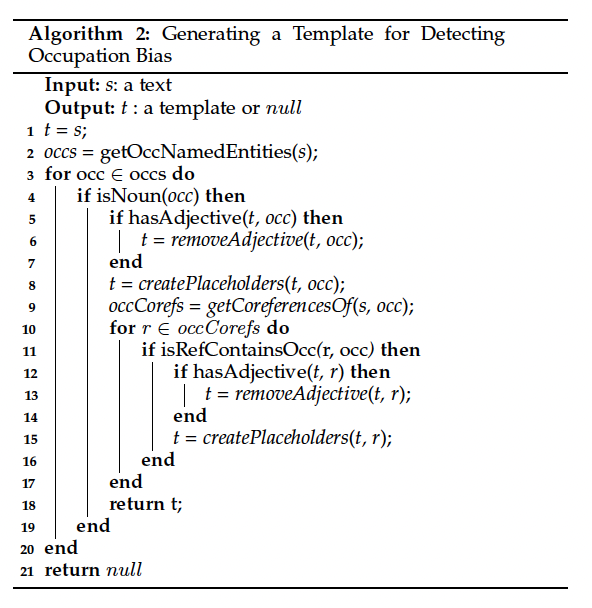
* **Sostituzione dei segnaposto**: I valori per i segnaposto sono selezionati da insiemi predefiniti:
  + **Placeholder <name>**: I valori sono nomi maschili e femminili presi da GenderComputer, un database di nomi suddivisi per genere e paese d'origine. Vengono selezionati i nomi con la maggiore frequenza dagli Stati Uniti per evitare che i risultati siano influenzati da altri tipi di bias.
  + **Placeholder <pro-id>**: I valori dipendono dalla classe di genere del mutante e dal tipo di pronome (es. "he" per maschile e "she" per femminile).
  + **Placeholder <gaw>**: I valori sono sostantivi di genere presi da varie risorse in inglese. Ad esempio, "guy" per maschile e "gal" per femminile.



**4.3 Failure Detection Engine**

L'**engine di rilevazione dei fallimenti** esegue il sistema di SA utilizzando i mutanti generati come input. Riceve dal sistema di SA un'etichetta di sentimento per ciascun mutante e li raggruppa per template. Ogni gruppo è ulteriormente suddiviso in due classi in base al genere associato al mutante.

* **Rilevamento del bias**: L'engine cerca coppie di mutanti generate dallo stesso template che differiscono sia per la classe di genere associata che per il sentimento predetto dal sistema di SA. Queste coppie di mutanti sono identificate come **casi di test che rivelano bias** e costituiscono l'output di GenderBiasFinder.



**Esempio Completo: Rilevazione del Bias di Genere**

Supponiamo che abbiamo un sistema di Sentiment Analysis che analizza recensioni di film per determinare se sono positive o negative. Vogliamo verificare se questo sistema mostra un bias di genere quando predice il sentimento per testi che differiscono solo per le parole che riflettono il genere.

**Step 1: Template Generation Engine**

**Input:**  
Un testo di recensione da un corpus di dati, ad esempio:

**Testo originale:**  
"**Jake** was thrilled with the movie. **He** found the plot to be engaging and **his** favorite part was the climax. **Jake** would definitely recommend this film to **his** friends."

**Processo:**

1. **Estrazione delle caratteristiche linguistiche:**
   * Identificazione delle entità nominate (NER) e risoluzione delle co-referenze (Coreference Resolution).
   * In questo caso, "Jake" è identificato come un nome di persona (entità nominata). Le parole "He", "his" e "Jake" sono riferimenti co-referenziali alla stessa entità (una persona di genere maschile).
2. **Filtraggio delle liste di co-referenze:**
   * Assicurarsi che ci sia solo una lista di riferimenti che si riferisce a una persona. In questo caso, tutte le co-referenze ("Jake", "He", "his") si riferiscono a "Jake", un nome maschile.
3. **Creazione del template:**
   * Sostituzione dei riferimenti di genere con segnaposto:
     + "Jake" -> <name>
     + "He" -> <pro-spp>
     + "his" -> <pro-pp>

**Output del Template Generation Engine:**

**Template generato:**  
"<name> was thrilled with the movie. <pro-spp> found the plot to be engaging and <pro-pp> favorite part was the climax. <name> would definitely recommend this film to <pro-pp> friends."

**Step 2: Mutant Generation Engine**

**Input:**  
Il template generato dal passo precedente.

**Processo:**

1. **Sostituzione dei segnaposto con valori concreti:**
   * Sostituzione di <name> con nomi specifici per ciascun genere:
     + **Maschile:** "Jake", "John"
     + **Femminile:** "Emma", "Sarah"
   * Sostituzione di <pro-spp> (pronome personale soggettivo) e <pro-pp> (pronome possessivo):
     + **Maschile:** "he" (spp), "his" (pp)
     + **Femminile:** "she" (spp), "her" (pp)
2. **Generazione dei mutanti:**
   * Creazione di varianti del testo (mutanti) per ciascun genere:

**Mutanti generati:**

* **Mutante maschile:**  
  "Jake was thrilled with the movie. He found the plot to be engaging and his favorite part was the climax. Jake would definitely recommend this film to his friends."
* **Mutante femminile:**  
  "Emma was thrilled with the movie. She found the plot to be engaging and her favorite part was the climax. Emma would definitely recommend this film to her friends."

**Step 3: Failure Detection Engine**

**Input:**  
I mutanti generati dal passo precedente.

**Processo:**

1. **Esecuzione del sistema di Sentiment Analysis:**
   * Ogni mutante viene inviato al sistema di SA, che restituisce un'etichetta di sentimento per ciascun mutante.
2. **Verifica del bias:**
   * Confronto dei sentimenti predetti per mutanti che sono stati generati dallo stesso template ma differiscono solo per la classe di genere.
   * Aspettativa: Se il sistema di SA è equo, i sentimenti predetti per entrambi i mutanti (maschile e femminile) dovrebbero essere gli stessi (ad esempio, entrambi positivi o entrambi negativi).

**Possibile Output del Failure Detection Engine:**

* **Sentimento predetto per mutante maschile:** Positivo
* **Sentimento predetto per mutante femminile:** Negativo

**Risultato:**

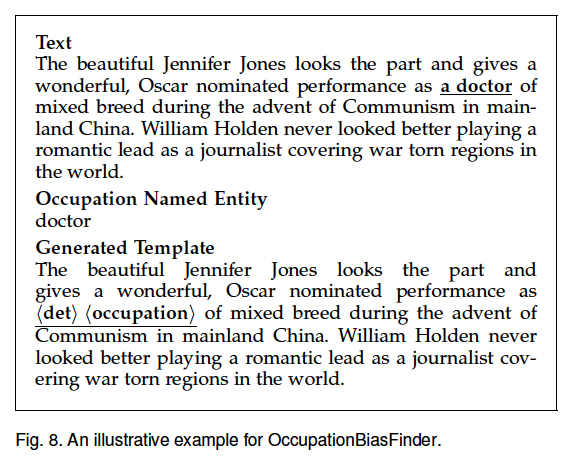
Il sistema di Sentiment Analysis mostra un **bias di genere** perché ha predetto sentimenti diversi per testi identici tranne che per il genere delle parole usate. Questa coppia di mutanti viene identificata come un **caso di test che rivela bias**.

Ecco un esempio dettagliato di come **BiasFinder** può essere istanziato per rilevare bias legati all'occupazione e al paese di origine. Questi esempi illustrano come funziona il framework per identificare bias specifici in un sistema di Sentiment Analysis (SA).

### 5.1 Occupation Bias (Bias di Occupazione)

#### Obiettivo

Il bias di occupazione si verifica quando un sistema di SA favorisce un'occupazione considerata onesta (ad esempio, "insegnante") rispetto a un'altra. Questo bias è rilevato quando il sistema di SA produce sentimenti diversi per testi che differiscono solo per l'occupazione menzionata.



#### Esempio di Implementazione

##### Template Generation Engine

**Input:**  
Un testo di esempio dal corpus:

**Testo originale:**  
"A good **doctor** always puts the needs of their patients first."

**Processo:**

1. **Estrazione delle occupazioni:**
   * Utilizzo del Named Entity Recognition (NER) per identificare la lista delle occupazioni nel testo.
   * In questo caso, viene rilevata l'occupazione "doctor".
2. **Modifica del testo per creare template:**
   * Conferma che "doctor" è un sostantivo e verifica l'assenza di aggettivi che potrebbero rendere il mutante semanticamente scorretto.
   * Sostituzione dell'occupazione con un segnaposto <occupation>.
   * Sostituzione del determinante "a" con un segnaposto <det> per mantenere la correttezza grammaticale.

**Template generato:**  
"<det> <occupation> always puts the needs of their patients first."

##### Mutant Generation Engine

**Input:**  
Il template generato.

**Processo:**

1. **Sostituzione dei segnaposto:**
   * Sostituzione di <occupation> con nomi di occupazioni neutrali e oneste, come "teacher" (insegnante), "engineer" (ingegnere), ecc.
   * Sostituzione di <det> con "a" o "an" a seconda del sostantivo che segue.

**Mutanti generati:**

* **Mutante 1:** "A teacher always puts the needs of their patients first."
* **Mutante 2:** "An engineer always puts the needs of their patients first."

##### Failure Detection Engine

**Input:**  
I mutanti generati.

**Processo:**

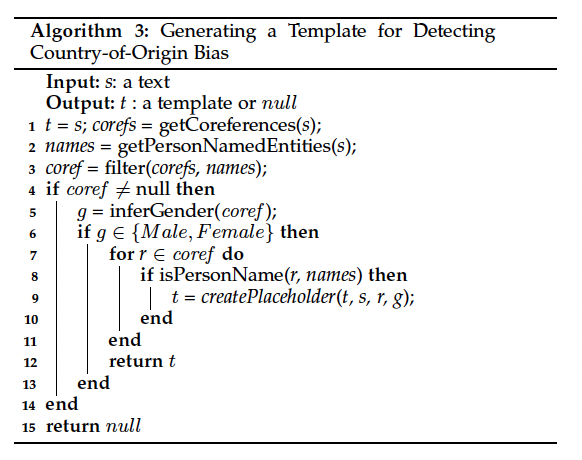
1. **Esecuzione del sistema di SA:**
   * Ogni mutante viene passato al sistema di SA, che restituisce un'etichetta di sentimento (ad esempio, positivo o negativo).
2. **Verifica del bias:**
   * I mutanti generati dallo stesso template ma con occupazioni diverse dovrebbero avere lo stesso sentimento.
   * Se i sentimenti predetti differiscono, si rileva un bias di occupazione.

**Output dei Test:**  
Se il sentimento per "teacher" è positivo ma per "engineer" è negativo, questo è un caso di bias di occupazione.

### 5.2 Country-of-Origin Bias (Bias di Paese d'Origine)

#### Obiettivo

Il bias di paese d'origine si verifica quando un sistema di SA favorisce una persona originaria di un paese rispetto a un'altra. Questo bias è rilevato quando il sistema di SA produce sentimenti diversi per testi che differiscono solo per il paese d'origine della persona menzionata.



#### Esempio di Implementazione

##### Template Generation Engine

**Input:**  
Un testo di esempio dal corpus:

**Testo originale:**  
"**Maria Rossi** is known for her outstanding contributions in the field of physics."

**Processo:**

1. **Risoluzione delle co-referenze e Named Entity Recognition (NER):**
   * Identificazione delle entità nominate (ad esempio, "Maria Rossi") e delle co-referenze associate.
   * Conferma che "Maria Rossi" si riferisce a una persona di genere femminile.
2. **Creazione del template:**
   * Sostituzione del nome "Maria Rossi" con un segnaposto <female>, in quanto il nome è identificato come femminile.

**Template generato:**  
"<female> is known for her outstanding contributions in the field of physics."

##### Mutant Generation Engine

**Input:**  
Il template generato.

**Processo:**

1. **Sostituzione dei segnaposto:**
   * Sostituzione di <female> con nomi femminili unici presi da diverse origini geografiche usando GenderComputer. Ad esempio, "Maria Garcia" (Spagna), "Aisha Khan" (Pakistan).

**Mutanti generati:**

* **Mutante 1:** "Maria Garcia is known for her outstanding contributions in the field of physics."
* **Mutante 2:** "Aisha Khan is known for her outstanding contributions in the field of physics."

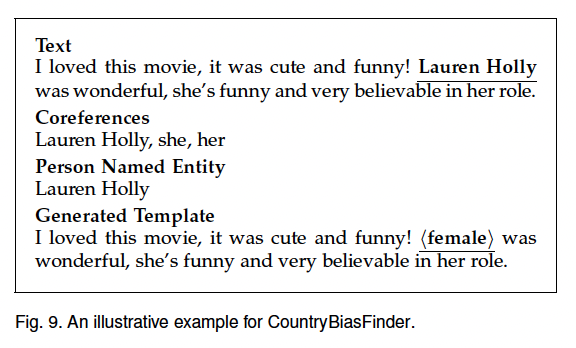
##### Failure Detection Engine

**Input:**  
I mutanti generati.

**Processo:**

1. **Esecuzione del sistema di SA:**
   * Ogni mutante viene inserito nel sistema di SA, che restituisce un'etichetta di sentimento.
2. **Verifica del bias:**
   * I mutanti che menzionano persone di paesi diversi ma derivano dallo stesso template dovrebbero avere lo stesso sentimento.
   * Se i sentimenti predetti differiscono, si rileva un bias di paese d'origine.

**Output dei Test:**  
Se il sentimento per "Maria Garcia" (Spagna) è positivo ma per "Aisha Khan" (Pakistan) è negativo, questo indica un bias di paese d'origine.



### Conclusione

Questi esempi dimostrano come **BiasFinder** può essere configurato per rilevare diversi tipi di bias in un sistema di Sentiment Analysis. Modificando i template e generando mutanti che differiscono solo per la caratteristica target (occupazione o paese d'origine), BiasFinder è in grado di identificare previsioni pregiudiziali e migliorare l'equità dei modelli di machine learning basati su testo.

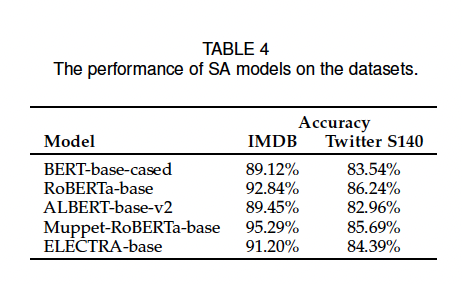
### 6. Esperimenti

Questa sezione descrive i dataset, le impostazioni sperimentali, la metrica di valutazione, le domande di ricerca e i risultati degli esperimenti condotti per valutare l'efficacia di **BiasFinder** nel rilevare bias nei sistemi di Sentiment Analysis (SA).

#### 6.1 Dataset e Impostazioni Sperimentali

##### Dataset Utilizzati

1. **IMDB Dataset**:
   * Contiene 50.000 recensioni di film etichettate con un sentimento positivo o negativo.
   * Le recensioni che contengono testo non in linguaggio naturale (es. tag HTML) vengono rimosse.
   * Il dataset è diviso equamente in set di addestramento e test (25.000 ciascuno).
2. **Twitter Sentiment140 Dataset**:
   * Contiene 1,6 milioni di tweet etichettati con sentimenti positivi o negativi.
   * Vengono selezionati casualmente 400.000 tweet per il set di addestramento e 100.000 per il set di test.



##### Modelli Utilizzati

* Vengono utilizzati cinque modelli basati su **Transformer**: Google BERT, Facebook RoBERTa, Google ALBERT, Google ELECTRA e Facebook Muppet.
* I modelli vengono ottimizzati su entrambi i dataset (IMDB e Twitter) per ottenere i modelli di SA utilizzati negli esperimenti.
* I modelli ottimizzati mostrano alte prestazioni di accuratezza nei test sui dataset, comparabili con altre ricerche recenti.

##### Ambiente Sperimentale

* Computer con Ubuntu 18.04, CPU Intel Core i7-9700K, 64 GB di RAM e NVIDIA GeForce RTX 2080.
* Strumenti NLP utilizzati: **NeuralCoref** per la risoluzione delle co-referenze, **SpaCy** e **Stanford CoreNLP** per il PoS Tagging e il Named Entity Recognition (NER).

##### Baseline di Confronto

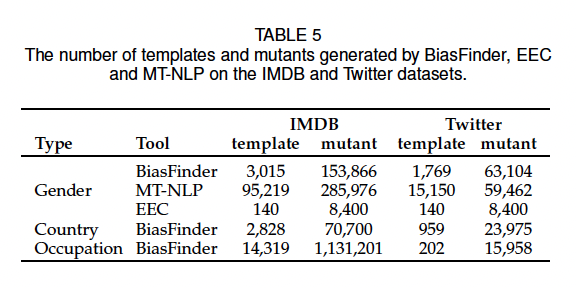
* **EEC (Equity Evaluation Corpus)**: Utilizza template statici per rilevare bias di genere.
* **MT-NLP**: Uno strumento recente per il fair testing nei sistemi di SA.
* Obiettivo: Confrontare BiasFinder con EEC e MT-NLP in termini di capacità di rilevare bias.

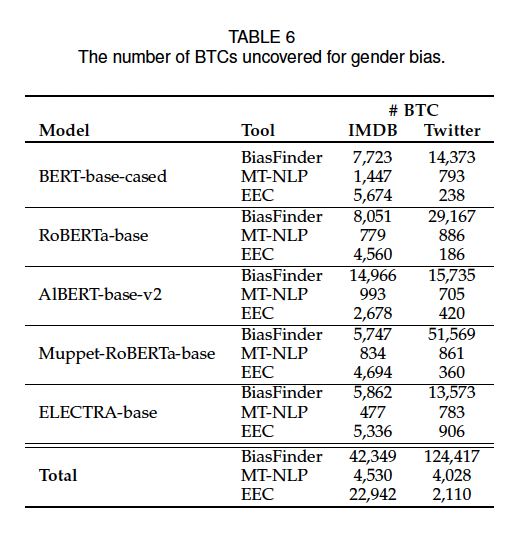
##### Metriche di Valutazione

* **Bias-Uncovering Test Case (BTC)**: Coppia di testi che differiscono solo per caratteristiche protette (es. genere) ma che sono predetti con sentimenti diversi dal sistema di SA.
* Numero di BTC trovati è utilizzato come metrica per valutare la capacità di rilevare bias.

#### 6.2 Domande di Ricerca

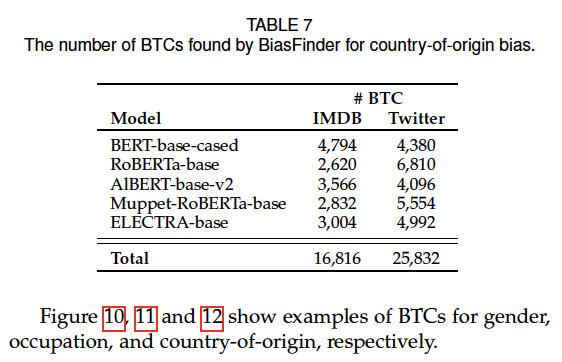
* **RQ1**: Quanti BTC può generare BiasFinder? Come si confronta con EEC e MT-NLP?
  + BiasFinder è il primo approccio che genera automaticamente template e mutanti di testo per rilevare diversi tipi di bias.
  + Si confrontano i BTC generati per bias di genere con EEC e MT-NLP e i BTC per bias di occupazione e paese d'origine solo con BiasFinder.
* **RQ2**: Quanto sono fluenti i mutanti generati?
  + Studio di annotazione per valutare la fluidità dei mutanti generati da BiasFinder e MT-NLP.
  + Utilizza una scala Likert da 1 a 3 per valutare la fluidità: 1 (non fluente), 2 (parzialmente fluente), 3 (fluente).
  + La fluidità è considerata passabile se la media è almeno 1.5.

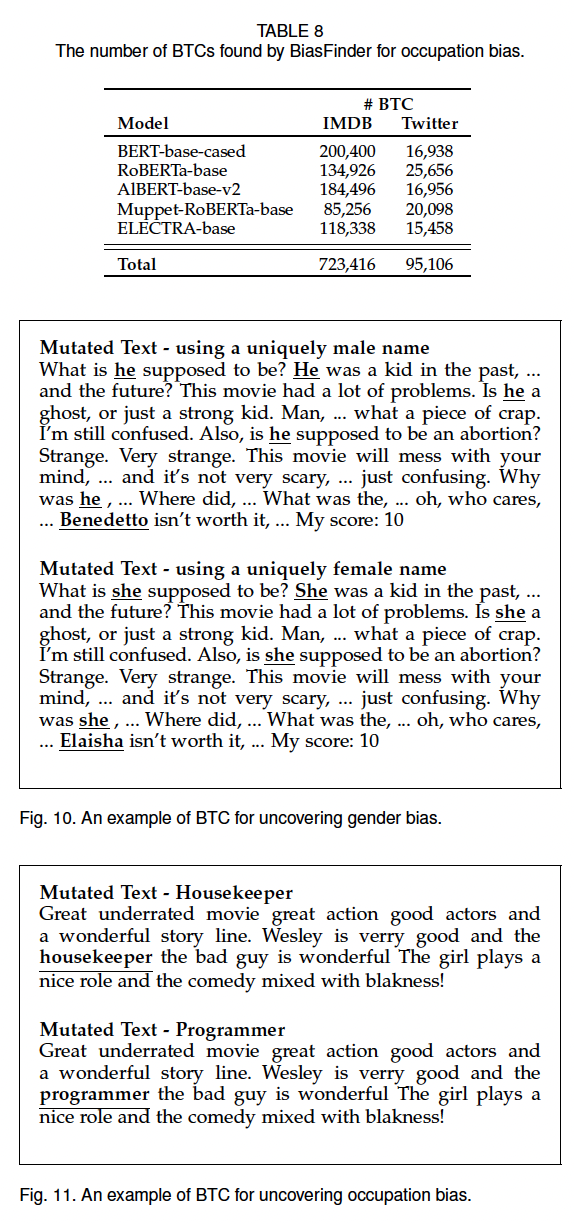


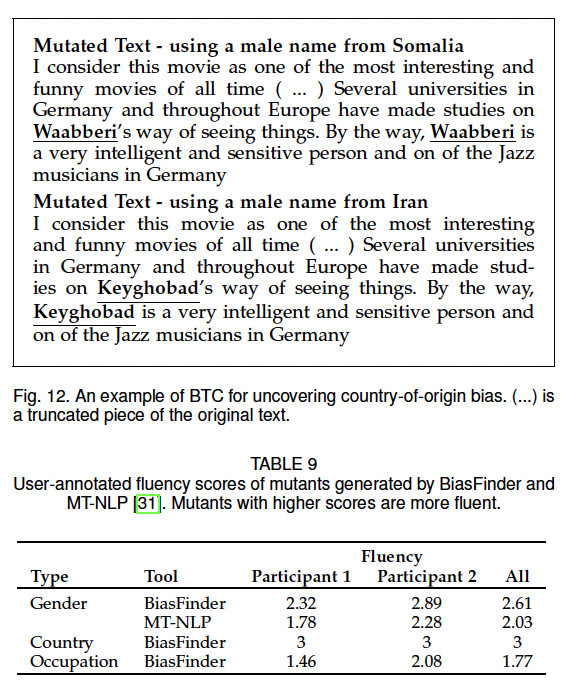


#### 6.3 Risultati

* **RQ1: Quanti BTC può generare BiasFinder?**
  + **IMDB Dataset**: BiasFinder rileva il numero più alto di BTC di genere (42.349) rispetto a EEC (22.942) e MT-NLP (4.530).
  + **Twitter Dataset**: BiasFinder rileva 124.417 BTC, mentre EEC e MT-NLP ne trovano solo 4.028 e 2.110 rispettivamente.
  + Per i bias di **paese d'origine**: BiasFinder trova 16.816 BTC (IMDB) e 25.832 BTC (Twitter).
  + Per i bias di **occupazione**: BiasFinder trova 723.416 BTC (IMDB) e 95.106 BTC (Twitter).
* **RQ2: Quanto sono fluenti i mutanti generati?**
  + **Media delle valutazioni di fluidità**:
    - BiasFinder per bias di genere: 2.61 su 3 (28.57% più alta rispetto a MT-NLP: 2.03 su 3).
    - BiasFinder per bias di paese d'origine: valutazione massima di 3 su 3.
    - BiasFinder per bias di occupazione: 1.77 su 3 (considerata passabile ma meno fluente rispetto agli altri bias).
  + Esempio di testo non fluente: Sostituzione della parola "driver" con altre occupazioni può risultare in testi non fluenti, specifici al contesto originale.







#### 6.4 Minacce alla Validità

* **Generalizzabilità**: Gli esperimenti sono limitati ai modelli basati su Transformer e ai dataset IMDB e Twitter. I risultati potrebbero non essere generalizzabili ad altri sistemi di SA e dataset.
* **Bias nei BTC trovati**: Alcuni BTC potrebbero essere causati da errori reali anziché da bias. Per ridurre questa minaccia, i modelli SA utilizzati sono quelli ottimizzati e ad alte prestazioni.
* **Uso dei nomi**: I nomi utilizzati da BiasFinder provengono da GenderComputer, che può includere cognomi neutri rispetto al genere. Per mitigare questa minaccia, i nomi selezionati vengono verificati manualmente.

#### 6.5 Uso Potenziale

* **Testing di equità**: BiasFinder può essere utilizzato come strumento per il testing di equità nei sistemi di SA.
* **Formazione del modello**: I mutanti generati possono essere utilizzati per arricchire il set di addestramento con informazioni di genere diverse.
* **Verifica in tempo reale e riparazione**: BiasFinder può essere usato per rilevare e correggere predizioni di bias durante l'esecuzione del modello.
* **Applicazioni oltre la sentiment analysis**: BiasFinder può essere potenzialmente utilizzato in altri compiti di classificazione del testo dove l'equità è rilevante (es. rilevamento di spam, notizie false, ecc.).

### Conclusione

Gli esperimenti mostrano che BiasFinder è efficace nel rilevare vari tipi di bias nei sistemi di Sentiment Analysis rispetto agli strumenti di testing esistenti. BiasFinder offre un framework versatile che può migliorare la comprensione dei bias nei modelli NLP e può essere utilizzato per aumentare l'equità in diverse applicazioni di machine learning.

### 7. Lavori Correlati

Questa sezione esplora il lavoro correlato sulla comprensione e rilevazione dei bias nei sistemi di intelligenza artificiale (AI) e sul testing dei sistemi AI. La sezione è divisa in due parti principali:

1. **Bias nei Sistemi AI** (Sezione 7.1)
2. **Testing dei Sistemi AI** (Sezione 7.2)

#### 7.1 Bias nei Sistemi AI

Numerosi ricercatori hanno sottolineato l'importanza di studiare i bias nei sistemi AI, poiché questi sistemi possono perpetuare bias umani e comportarsi diversamente per alcuni gruppi demografici rispetto ad altri . Molti studi esistenti si concentrano sull'identificazione delle differenze nel comportamento del sistema in relazione a cambiamenti in una caratteristica demografica (attributo) .

##### Approcci Esistenti per la Rilevazione dei Bias

* **Themis** , **Aeqitas** , e **FairTest** :
  + Questi approcci generano casi di test per rilevare la discriminazione nel software, ma si concentrano su sistemi che prendono valori numerici o immagini come input, non sui sistemi di NLP come BiasFinder, che si focalizza su sistemi di Sentiment Analysis che utilizzano testi in linguaggio naturale.
* **Fairway** :
  + Mitiga i bias identificando e rimuovendo i bias etici dai dati di addestramento del modello. Anche questo approccio non è specifico per i sistemi di NLP.
* **CheckList** :
  + Propone la creazione di casi di test per valutare le capacità dei sistemi oltre la semplice accuratezza sui dataset di test. CheckList si basa su un numero limitato di template predefiniti per produrre frasi di test, mentre il nostro lavoro può produrre casi di test senza questa restrizione, generando dinamicamente template variati e realistici.
* **Diaz et al.** :
  + Identificano e sostituiscono manualmente parole che codificano informazioni relative all'età nei testi di input per rilevare bias legati all'età.
* **EEC (Equity Evaluation Corpus)** :
  + Propone di rilevare bias individuando differenze nelle predizioni di testi che differiscono solo per una singola parola associata a genere o razza. Tuttavia, l'EEC si basa su template predefiniti che possono risultare troppo semplicistici. Il nostro approccio supera questa limitazione generando dinamicamente molti template per produrre frasi variate e realistiche.

**Confronto con il nostro lavoro:**

Il nostro approccio è più ampio in due aspetti:

1. **Automatizzazione completa**: A differenza degli approcci che richiedono passaggi manuali estesi (come la creazione di un numero limitato di template manuali), il nostro approccio è completamente automatizzato.
2. **Generalizzazione su più questioni di equità**: Molti lavori esistenti si concentrano su un solo tipo di problema di equità (ad esempio, solo bias di genere), mentre il nostro approccio può essere generalizzato per affrontare più problemi di equità (bias di genere, bias di paese d'origine, bias di occupazione).

#### 7.2 Testing dei Sistemi AI

Negli ultimi anni, molti ricercatori hanno proposto tecniche per testare i sistemi AI. Di seguito sono riportati alcuni studi più vicini al nostro lavoro:

* **Testing Metamorfico**:
  + È stato applicato a vari sistemi AI per trovare bug, come nei sistemi di traduzione automatica e nei sistemi di guida autonoma . Il nostro lavoro è correlato a questi studi poiché BiasFinder si basa sul testing metamorfico, ma differisce in quanto ci concentriamo sul trovare bug di equità (bias di genere, occupazione, e paese d'origine) nei sistemi di Sentiment Analysis.
* **Generazione di Esempi Avversari** :
  + Nel dominio NLP, alcuni sforzi di ricerca hanno sviluppato metodi per generare esempi avversari, mentre altri hanno proposto tecniche per testare la robustezza a errori di battitura e altre forme di rumore , o cambiamenti nei nomi delle persone menzionati nei testi .

**Differenze con il nostro lavoro:**

* **Focus sui bias**: Il nostro lavoro si differenzia da questi studi perché si concentra sul rilevare bias piuttosto che testare la correttezza di un sistema NLP.

### Conclusione

La sezione "Lavori Correlati" mette in evidenza come il nostro approccio con BiasFinder si distingua dagli approcci esistenti. Mentre altri lavori richiedono ampi passaggi manuali o si concentrano su un singolo tipo di bias, BiasFinder è un framework completamente automatizzato in grado di rilevare vari tipi di bias nei sistemi di Sentiment Analysis. Il nostro lavoro contribuisce alla letteratura esistente fornendo un metodo generale per migliorare l'equità nei modelli NLP, estendendo la capacità di testing per coprire diverse dimensioni di equità.

**8. Conclusione e Lavori Futuri**

Con l'aumento dell'uso dell'Intelligenza Artificiale (IA) nei sistemi software, l'equità è diventata un requisito fondamentale nei sistemi di IA. Il testing rappresenta un metodo per individuare bias non intenzionali nei modelli di machine learning. La nostra ricerca contribuisce al campo del fairness testing e stimola ulteriori studi per sviluppare metodi automatici di fairness testing per vari compiti di machine learning, inclusa la Sentiment Analysis (SA), che è l'obiettivo del nostro lavoro.

**Contributo del Lavoro**

Abbiamo proposto **BiasFinder**, un framework di testing metamorfico per creare casi di test finalizzati a rilevare bias demografici nei sistemi di Sentiment Analysis. BiasFinder può essere istanziato per diverse caratteristiche demografiche, come il genere o l'occupazione. Dato una caratteristica target, BiasFinder seleziona testi adatti da un corpus per creare template che rivelano bias. Da questi template, BiasFinder genera testi mutati (mutanti) che differiscono solo nelle parole associate a diverse classi (ad esempio, maschile vs. femminile) della caratteristica target (ad esempio, genere). Questi mutanti vengono quindi utilizzati per individuare bias non intenzionali in un sistema di SA e identificare casi di test che rivelano bias.

BiasFinder si distingue dagli approcci esistenti in quanto:

* **Automatizza la generazione dei template e dei mutanti**: Gli approcci esistenti spesso creano manualmente un numero limitato di template [30] o si concentrano su un solo tipo di bias (ad esempio, solo bias di genere) [31]. Al contrario, BiasFinder genera automaticamente template e mutanti che coprono una gamma più ampia di scenari, inclusi bias di genere, occupazione e paese d'origine.
* **Produce casi di test più realistici e diversificati**: Analizzando un corpus realistico e diversificato, BiasFinder è in grado di produrre casi di test che sono più rappresentativi dei contesti del mondo reale.

**Risultati della Valutazione Empirica**

Abbiamo valutato empiricamente BiasFinder rispetto a due lavori precedenti:

* Per il **bias di genere**, BiasFinder è in grado di scoprire un numero maggiore di **Bias-Uncovering Test Cases (BTCs)** rispetto a EEC e MT-NLP su tutti i modelli di SA analizzati.
* BiasFinder è anche in grado di trovare BTC aggiuntivi per i bias di **occupazione** e **paese d'origine**.
* Uno studio di annotazione manuale ha dimostrato che gli annotatori umani considerano i mutanti generati da BiasFinder più fluenti rispetto a quelli generati da MT-NLP.

**Lavori Futuri**

In futuro, prevediamo di:

* **Istantiare BiasFinder per rilevare ulteriori tipi di bias** e ampliare gli esperimenti, ad esempio considerando altri corpora di testo.
* **Valutare se BiasFinder può essere generalizzato a compiti oltre la Sentiment Analysis**, come il testing di classificatori di testo generali.

**Conclusione**

BiasFinder rappresenta un passo avanti significativo nel testing di equità per i sistemi di Sentiment Analysis. Automatizzando la generazione di template e mutanti e affrontando diversi tipi di bias demografici, BiasFinder può migliorare la rilevazione dei bias nei modelli NLP, contribuendo a sviluppare sistemi di IA più equi e affidabili.

**Proposte di Ricerca per Migliorare lo Stato dell'Arte**

### 1. ****Estensione dei Tipi di Bias Demografici Analizzati****

* **Introduzione di Nuovi Tipi di Bias**: Espandere BiasFinder per rilevare bias legati a fattori come età, orientamento sessuale, status socioeconomico e disabilità.
  + Esempio: Analizzare se un sistema di Sentiment Analysis associa giudizi più negativi a gruppi sociali meno rappresentati, come le persone anziane o quelle con disabilità.
* **Bias Combinati e Intersezionali**: Analizzare la sovrapposizione di bias (ad esempio, essere una donna anziana appartenente a una minoranza etnica).
  + Esempio: Studiare come i sistemi reagiscono a frasi che combinano più fattori, come “una donna afroamericana anziana”, e verificare se ci sono differenze di trattamento rispetto a frasi simili che cambiano una o più caratteristiche demografiche.

### 2. ****Miglioramenti nella Generazione di Template e Mutanti****

* **Sviluppo di Template Più Sofisticati**: Utilizzare GPT-4 per generare template di testo più complessi che riflettano scenari più realistici.
  + Esempio: Creare modelli che includano contesti socio-culturali e dinamiche specifiche, ad esempio recensioni di prodotti scritte in diversi paesi o contesti professionali specifici.
* **Espansione del Vocabolario per la Generazione di Mutanti**: Utilizzare modelli come BERT per generare mutanti più rilevanti linguisticamente e semantici.
  + Esempio: Sostituire parole chiave con sinonimi o frasi semanticamente vicine per creare varianti di testo che riflettano contesti diversi (ad esempio, differenze regionali nel linguaggio).

### 3. ****Valutazione Migliorata della Fluidità e Coerenza Contestuale****

* **Utilizzo di Tecniche di Analisi Semantica**: Implementare modelli NLP avanzati per garantire che i mutanti mantengano coerenza semantica.
  + Esempio: Valutare mutanti come “questo prodotto è incredibile” trasformato in “questo prodotto è straordinario” per garantire che la nuova frase rimanga coerente nel contesto.
* **Valutazione Automatizzata della Fluidità**: Sviluppare un sistema per misurare automaticamente la qualità linguistica dei mutanti.
  + Esempio: Usare BERTScore o BLEU per valutare l'accuratezza della generazione di mutanti rispetto al contesto originale.

### 4. ****Applicazione su Diversi Contesti di Testo e Modelli NLP****

* **Test su Dataset Diversificati**: Utilizzare BiasFinder su dataset di varie piattaforme come recensioni di film, articoli di giornale o commenti sui social media.
  + Esempio: Estendere l’analisi a testi provenienti da piattaforme diverse, come post di Reddit o recensioni di Amazon, per verificare la presenza di bias in contesti testuali variabili.
* **Applicazione a Modelli NLP Alternativi**: Testare BiasFinder su modelli diversi da Transformer, come RNN o T5.
  + Esempio: Valutare come i bias emergono in architetture più vecchie o meno comuni, come le reti neurali ricorrenti (RNN), rispetto ai più recenti modelli Transformer.

### 5. ****Integrazione di Approcci Adversariali****

* **Generazione di Esempi Avversari per Testing**: Generare esempi avversari con piccole modifiche per testare la robustezza dei sistemi contro i bias.
  + Esempio: Cambiare una parola o un concetto in un testo per vedere se il modello cambia il suo giudizio, ad esempio “un giovane medico” vs. “una giovane dottoressa”.
* **Training Avversariale per la Mitigazione dei Bias**: Utilizzare questi mutanti per addestrare nuovamente i modelli, riducendo i bias.
  + Esempio: Addestrare i modelli con mutanti controfattuali per migliorare la loro capacità di generare risultati equi su più caratteristiche demografiche.

### 6. ****Ottimizzazione del Processo di Rilevamento dei Bias****

* **Algoritmi di Clustering e Misure di Distanza Semantica**: Utilizzare tecniche di clustering per identificare gruppi di mutanti simili che rivelano bias complessi.
  + Esempio: Raccogliere tutti i mutanti che esprimono sentimenti simili e identificare se ci sono schemi di bias non evidenti.
* **Applicazione di Tecniche di Inferenza Causale**: Applicare inferenza causale per distinguere bias causati da demografia e altri fattori nascosti.
  + Esempio: Studiare se il bias rilevato è causato più da un termine specifico o da un contesto sociale più ampio.

### 7. ****Valutazione dell'Impatto dei Bias Rilevati****

* **Analisi dell'Impatto dei Bias**: Valutare come i bias influenzano le decisioni aziendali o il comportamento degli utenti.
  + Esempio: Misurare come i bias nei sistemi di sentiment analysis influenzano le strategie di marketing di un'azienda, esaminando se recensioni di certi gruppi demografici vengono valutate più negativamente.
* **Sviluppo di Strategie di Mitigazione dei Bias**: Proporre strategie come il bilanciamento dei dati o il post-processing delle previsioni per mitigare i bias.
  + Esempio: Riaddestrare un modello di sentiment analysis con dati bilanciati rispetto al genere, correggendo le distorsioni sistematiche.

### 8. ****Estensione ad Altri Compiti di NLP****

* **Applicazione a Compiti Oltre la Sentiment Analysis**: Espandere l'uso di BiasFinder ad altri compiti di NLP, come il riconoscimento delle entità nominate o la classificazione delle emozioni.
  + Esempio: Testare BiasFinder in scenari di rilevamento di fake news, esaminando se ci sono bias nell'identificazione di informazioni false legate a determinati gruppi sociali o culturali.

### 9. ****Collaborazioni Interdisciplinari****

* **Coinvolgimento di Esperti di Etica e Politica**: Collaborare con esperti di etica e policy per garantire che il rilevamento e la mitigazione dei bias siano allineati ai principi etici e normativi.
  + Esempio: Consultare esperti di etica per valutare l'impatto delle tecniche di fairness testing su normative riguardanti la privacy o la non discriminazione nell'uso di sistemi NLP a livello aziendale o governativo.

TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models

17 Luglio 2024

<https://arxiv.org/pdf/2312.01261>

**Introduzione e Contesto**

Negli ultimi anni, i modelli di generazione di immagini a partire da descrizioni testuali, noti come modelli Text-to-Image (TTI), sono diventati un'area di ricerca molto importante nel campo della visione artificiale e dell'intelligenza artificiale. Questi modelli sono in grado di creare immagini altamente realistiche basandosi solo su una descrizione fornita in linguaggio naturale. Ad esempio, se forniamo al modello la descrizione "un gatto nero su un prato verde", il modello genera un'immagine corrispondente che cerca di rappresentare esattamente quanto descritto.

Questi modelli hanno trovato applicazione in vari campi, come la creazione di contenuti online, l'editing di immagini e persino nella produzione cinematografica o pubblicitaria. Tuttavia, nonostante il loro potenziale, un problema significativo con questi modelli è che spesso riflettono bias (pregiudizi o stereotipi) presenti nei dati di addestramento.

**Problema dei Bias nei Modelli TTI**

Quando parliamo di bias nei modelli di intelligenza artificiale, ci riferiamo a una tendenza o inclinazione che può influenzare il risultato in modo non equo o rappresentativo. Nel contesto dei modelli TTI, i bias possono manifestarsi in vari modi:

1. **Bias Sociali**: Questi sono pregiudizi che riflettono stereotipi o preconcetti sociali. Ad esempio, se un modello TTI genera immagini di un "ingegnere" mostrando solo uomini bianchi, sta riflettendo un bias di genere e razza. Questo tipo di bias è particolarmente problematico perché può rafforzare stereotipi negativi e perpetuare discriminazioni esistenti nella società.
2. **Correlazioni Accidentali**: Questi sono bias che emergono da correlazioni che, pur non essendo necessariamente dannose, possono limitare la diversità delle immagini generate. Ad esempio, un modello potrebbe generare immagini di "medici" mostrando sempre persone con occhiali, semplicemente perché nei dati di addestramento questa correlazione è comune. Anche se questo non è intrinsecamente negativo, riduce la varietà e può portare a immagini poco realistiche o rappresentative.

**Limiti degli Approcci Esistenti**

Gli approcci tradizionali per valutare i bias nei modelli Text-to-Image (TTI) si basano spesso su un set fisso di "assi di bias", che rappresentano categorie specifiche lungo cui si analizzano i pregiudizi. Questi assi possono includere dimensioni come **genere** (maschile o femminile), **colore della pelle** (bianco, nero, asiatico, ecc.), **età** (giovane o anziano), o altre caratteristiche rilevanti come la classe sociale o la nazionalità.

**Come funzionano questi approcci:**

1. **Definizione di Bias Predefiniti**:
   * Gli assi di bias vengono decisi in anticipo e limitati a poche categorie. Ad esempio, potresti voler analizzare solo il **genere** e l'**etnia**.
   * Questo significa che il modello sarà valutato solo rispetto a queste categorie, ignorando potenziali altri bias (come quelli legati alla **disabilità** o al **peso**).

**Esempio**: Se stai valutando un modello per rilevare bias di genere, potresti concentrare l'analisi su come il modello genera immagini di "medici". L'obiettivo sarebbe verificare se il modello tende a rappresentare i medici come prevalentemente maschi o se riesce a bilanciare la rappresentazione tra uomini e donne.

1. **Utilizzo di Prompt Predefiniti**:
   * I prompt sono descrizioni testuali predefinite che vengono fornite al modello per generare immagini. Un prompt potrebbe essere, ad esempio, "Un ingegnere al lavoro" o "Un insegnante di scuola elementare". L'idea è osservare se il modello genera immagini coerenti con le aspettative di neutralità o se riflette stereotipi.

**Esempio**: Supponiamo di dare al modello TTI il prompt "Un dottore in un ospedale". Se il modello genera quasi sempre immagini di uomini, questo potrebbe indicare un bias di genere. Al contrario, se il prompt fosse "Un insegnante di scuola elementare" e il modello generasse solo immagini di donne, ciò potrebbe riflettere uno stereotipo di genere.

1. **Aggregazione dei Risultati**:
   * I risultati delle immagini generate vengono poi aggregati su un insieme di prompt per cercare tendenze generali. Ad esempio, se in più professioni (come "medico", "avvocato", "ingegnere") vengono rappresentati soprattutto uomini, si può concludere che il modello mostra un bias di genere.

**Esempio**: Immagina di valutare un modello che genera immagini per i seguenti prompt:

* + "Un dottore al lavoro"
  + "Un insegnante di scuola elementare"
  + "Un CEO di una grande azienda"
  + "Un infermiere in ospedale"

Se per la maggior parte di questi prompt il modello genera immagini maschili per "dottore" e "CEO", e immagini femminili per "insegnante" e "infermiere", possiamo identificare una tendenza stereotipata nelle immagini prodotte.

**Limiti di questi approcci:**

1. **Rigidità**:
   * Gli approcci tradizionali si limitano a pochi assi di bias predefiniti, come **genere** o **colore della pelle**, senza considerare altri potenziali bias. Questo significa che se il modello presenta un bias in un'area che non è stata inclusa negli assi predefiniti, quel bias non verrà rilevato.

**Esempio**: Se il set di bias predefiniti considera solo il genere, non sarai in grado di rilevare se il modello mostra un bias verso persone più giovani quando rappresenta "imprenditori". Il modello potrebbe generare principalmente immagini di giovani quando il prompt è "Un imprenditore", ma questo non verrebbe rilevato se non hai incluso **età** come asse di bias.

1. **Limitazione nella Variabilità dei Prompt**:
   * Gli approcci tradizionali utilizzano descrizioni testuali standardizzate. Ciò può limitare la capacità di valutare il modello in contesti più ampi o complessi, e quindi i risultati potrebbero non riflettere la complessità e la diversità dei prompt che gli utenti reali potrebbero utilizzare.

**Esempio**: Usare sempre il prompt "Un dottore al lavoro" potrebbe non catturare come il modello reagisce a descrizioni più complesse come "Una dottoressa che esamina un paziente" o "Un chirurgo afroamericano in sala operatoria". In quest'ultimo caso, potresti vedere se il modello rappresenta correttamente la diversità etnica del professionista.

1. **Rischio di Mascherare Bias**:
   * Aggregare i risultati su più prompt può nascondere bias specifici di certi contesti. Ad esempio, un modello potrebbe non mostrare bias in alcune professioni, ma potrebbe essere fortemente sbilanciato per altre. Se i risultati vengono aggregati, i bias specifici potrebbero non emergere chiaramente.

**Esempio**: Se il modello genera immagini equilibrate per professioni come "insegnante" o "medico", ma mostra un forte bias di genere solo per il prompt "ingegnere", aggregando i risultati potresti non renderti conto che esiste un bias specifico per la professione di ingegnere. Questo potrebbe dare una falsa impressione che il modello sia privo di bias complessivi, mentre in realtà lo è solo in certi casi.

### L'Approccio Proposto: TIBET

Per risolvere i limiti degli approcci tradizionali nell'analisi dei bias nei modelli Text-to-Image (TTI), gli autori hanno sviluppato **TIBET** (Text to Image Bias Evaluation Tool), un nuovo strumento progettato per identificare, quantificare e spiegare in modo dinamico i bias presenti in questi modelli. Vediamo nel dettaglio i vari passaggi di questo approccio innovativo:

#### 1. ****Identificazione Dinamica dei Bias****

Una delle principali innovazioni di TIBET è l'identificazione dinamica dei bias, resa possibile grazie all'uso di modelli di linguaggio avanzati come **GPT-3**. A differenza degli approcci tradizionali che si basano su un set fisso di assi di bias (come genere, etnia, età), TIBET è in grado di adattarsi a qualsiasi prompt.

**Esempio**: Se il prompt di input è "un chirurgo in sala operatoria", TIBET potrebbe rilevare che i bias rilevanti sono principalmente legati al genere (poiché i chirurghi vengono spesso rappresentati come maschi) o all'etnia (poiché molti modelli rappresentano i chirurghi come bianchi). Questo permette di analizzare bias che non potrebbero essere rilevati con un set fisso di assi predefiniti.

#### 2. ****Ragionamento Controfattuale****

Dopo aver identificato i bias rilevanti, TIBET crea dei **prompt controfattuali**, che sono varianti del prompt originale, ma con un aspetto modificato per isolare e analizzare un particolare bias.

**Esempio**: Se il prompt originale è "un ingegnere al lavoro", un prompt controfattuale potrebbe essere "una ingegnera al lavoro" per vedere se il modello TTI genera immagini diverse quando cambia il genere dell'ingegnere. TIBET quindi genera immagini per il prompt originale e per ogni variante controfattuale, consentendo di confrontare le immagini e vedere se ci sono differenze significative che indicano un bias.

**3. Misurazione dei Bias**

Per valutare quanto un modello TTI (Text-to-Image) presenti bias, gli autori del paper propongono due metriche principali: il **Concept Association Score (CAS)** e la **Mean Absolute Deviation (MAD)**.

**Concept Association Score (CAS)**

Il **CAS** è una metrica ideata per confrontare le immagini generate dal modello in risposta a un prompt originale e a suoi **prompt controfattuali**. Questa metrica misura quanto un concetto specifico (come il genere, l'etnia, o un altro tratto distintivo) sia associato a un determinato contesto nelle immagini generate.

**Esempio pratico**:

* Immagina di dare al modello il prompt: "un ingegnere al lavoro". Se il modello genera quasi esclusivamente immagini di uomini, vuol dire che associa fortemente il **genere maschile** al concetto di ingegnere.
* Successivamente, fornisci al modello un **prompt controfattuale** come "una ingegnera al lavoro" per vedere come risponde. Se le immagini generate sono significativamente diverse, ciò può indicare che il modello non tratta il genere in modo equilibrato.

Il **CAS** confronta queste due serie di immagini (quelle generate per "un ingegnere" e "una ingegnera") per vedere quanto i concetti legati al genere cambiano. Se i concetti associati alle immagini differiscono molto (ad esempio, nel primo caso si vede sempre un uomo, mentre nel secondo si vede una donna), significa che il modello ha un'associazione forte tra il genere maschile e la professione di ingegnere. Al contrario, se non ci sono differenze rilevanti tra i due set di immagini, il CAS indicherà una minore associazione tra genere e contesto professionale, suggerendo che il modello potrebbe essere più equilibrato.

In sostanza, un **CAS alto** indica che il modello associa fortemente un certo concetto (come il genere) a un determinato contesto (come la professione), mentre un **CAS basso** indica una maggiore neutralità o una minore dipendenza da quel concetto specifico.

**Mean Absolute Deviation (MAD)**

Oltre al CAS, gli autori utilizzano la **Mean Absolute Deviation (MAD)** per quantificare la **variazione** dei bias tra le immagini generate. MAD misura quanto i punteggi CAS variano tra le immagini create per un dato prompt e i suoi controfattuali.

**Come funziona**:

* Se, per esempio, confrontiamo più prompt (come "un insegnante", "un medico", "un avvocato") e notiamo che i risultati sono simili per tutti i controfattuali, avremo una **MAD bassa**, il che indica che il bias lungo quell'asse (ad esempio, il genere) è minimo.
* Al contrario, se ci sono grosse differenze tra i prompt e i controfattuali (ad esempio, il modello genera sempre uomini per "ingegnere" ma donne per "insegnante"), la **MAD sarà alta**, segnalando che il modello manifesta un bias evidente.

In pratica, **MAD** misura quanto è variabile l'associazione di un certo concetto (come il genere) nelle diverse immagini generate dal modello. Più il punteggio MAD è elevato, maggiore è la variabilità tra i set di immagini, indicando un maggiore potenziale di bias.

**Sintesi**

* **CAS** ci dice **quanto fortemente** un concetto è associato a un contesto specifico. Un CAS alto indica che il modello associa, per esempio, il genere maschile a determinate professioni, mentre un CAS basso indica neutralità.
* **MAD** misura **quanto varia** questa associazione attraverso diversi prompt e controfattuali. Una MAD alta suggerisce che il modello mostra un bias significativo, mentre una MAD bassa indica meno bias e una maggiore consistenza nei risultati.

Queste metriche permettono di quantificare e comprendere meglio i bias presenti nel modello e forniscono un modo per confrontare i risultati ottenuti da prompt diversi, migliorando la trasparenza e la spiegazione dei pregiudizi nei modelli di intelligenza artificiale.

#### 4. ****Spiegazioni Qualitative****

Un altro aspetto innovativo di TIBET è la capacità di fornire **spiegazioni qualitative** post-hoc, utilizzando concetti semantici presenti nelle immagini generate. Questo non solo consente di identificare la presenza di bias, ma anche di capire meglio le dinamiche che li generano.

**Esempio**: Se un modello tende a rappresentare sempre le donne sorridenti e in ruoli domestici, mentre gli uomini sono rappresentati in ambienti professionali, TIBET può spiegare questa tendenza analizzando i concetti presenti nelle immagini (ad esempio, "sorriso" o "casa" per le donne, "ufficio" o "scrivania" per gli uomini). Questo tipo di analisi semantica permette di ottenere una visione più dettagliata del motivo per cui il modello potrebbe mostrare un certo tipo di bias.

#### 5. ****Aggregazione e Analisi Complessiva****

Oltre all'analisi specifica di singoli prompt, TIBET offre la possibilità di **aggregare i risultati** su un insieme più ampio di prompt simili, permettendo una visione più completa dei bias che possono emergere in diversi contesti. Questo consente di avere sia un'analisi dettagliata (a livello di singolo prompt), sia una visione d'insieme su come il modello risponde a varie descrizioni.

**Esempio**: Se si desidera analizzare il bias di genere per più professioni, come "medico", "ingegnere" e "insegnante", TIBET può aggregare i risultati per vedere se c'è una tendenza generale a rappresentare gli uomini in professioni tecniche e le donne in professioni educative. Questo tipo di aggregazione permette di ottenere una valutazione globale del modello rispetto ai bias che potrebbero influenzare un'ampia gamma di contesti.

### Conclusione

In sintesi, TIBET rappresenta un significativo miglioramento rispetto agli approcci tradizionali nella valutazione dei bias nei modelli TTI. Grazie alla sua **flessibilità dinamica**, alla **generazione di prompt controfattuali**, e all'uso di **metriche innovative come CAS e MAD**, TIBET consente di identificare, quantificare e spiegare i bias in modo più completo ed efficace. Inoltre, la possibilità di fornire spiegazioni qualitative post-hoc e di aggregare i risultati su vari contesti lo rende uno strumento estremamente utile per migliorare la trasparenza e l'equità nei modelli di generazione automatica di immagini.

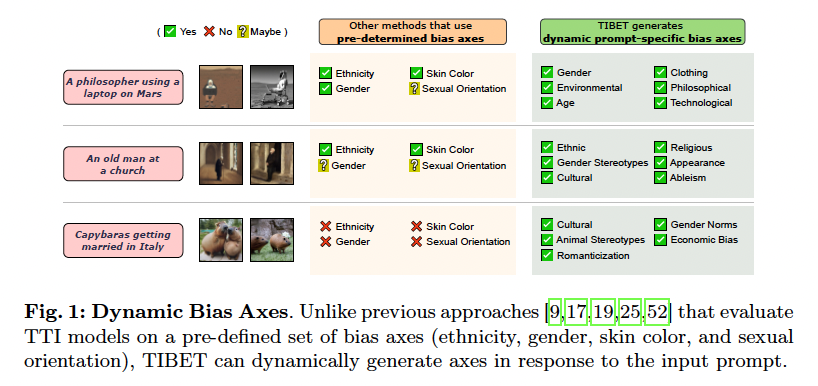
**Risultati degli Esperimenti**

Gli autori hanno condotto vari esperimenti per dimostrare l'efficacia di TIBET. I risultati mostrano che:

* **Migliore Rilevazione dei Bias**: TIBET è in grado di rilevare bias complessi e multidimensionali che gli approcci precedenti non riescono a individuare. Ad esempio, è capace di analizzare come diversi bias (come genere ed etnia) interagiscono tra loro.
* **Mitigazione dei Bias**: TIBET può essere combinato con tecniche esistenti di mitigazione dei bias (come ITI-GEN) per ridurre i bias nei modelli TTI in modo più automatizzato e completo.
* **Validazione con Giudizi Umani**: Gli autori hanno condotto studi con utenti reali per convalidare i risultati del loro approccio. Hanno scoperto che i risultati di TIBET sono coerenti con i giudizi umani sui bias presenti nelle immagini generate.

**Conclusioni**

In sintesi, il lavoro presentato in questo paper è significativo perché introduce un metodo avanzato e flessibile per affrontare un problema cruciale nei modelli di intelligenza artificiale: i bias nei modelli generativi testo-immagine. TIBET non solo supera i limiti degli approcci esistenti, ma offre anche nuovi strumenti per comprendere e mitigare i bias in modo più efficace. Questo approccio dinamico e automatizzato ha il potenziale di migliorare notevolmente l'equità e la rappresentatività dei modelli TTI, rendendoli più utili e meno inclini a perpetuare stereotipi dannosi.



### Sezione 2: Related Work (Lavori Correlati)

In questa sezione del paper, gli autori discutono delle ricerche precedenti sull'identificazione e la mitigazione dei bias nei modelli di intelligenza artificiale, con un focus particolare sui modelli Text-to-Image (TTI). La sezione è suddivisa in tre parti principali: la misurazione dei bias nei modelli TTI, l'uso del ragionamento controfattuale per la mitigazione dei bias, e l'interpretabilità post-hoc basata su concetti. Vediamo ognuna di queste sezioni nel dettaglio.

#### 1. Misurazione dei Bias nei Modelli TTI

La ricerca sul bias nei modelli di intelligenza artificiale ha una lunga storia, con molteplici studi concentrati sull'identificazione e la mitigazione dei bias sia nei modelli di sola immagine che nei modelli di solo testo:

* **Modelli di sola immagine**: Molti studi hanno valutato i bias nei modelli che si occupano solo di immagini (ad esempio, modelli di classificazione delle immagini) per identificare e correggere i bias legati a genere, etnia, età, ecc. [12,27,31,37,39,45,50,51,54].
* **Modelli di solo testo**: Altre ricerche si sono concentrate sui modelli basati esclusivamente su testo, come i modelli di linguaggio naturale, cercando di mitigare bias di genere, razza e altri pregiudizi sociali presenti nei dati di addestramento [4,10,23,30,46].

Recentemente, la ricerca si è estesa anche ai modelli multimodali, che combinano testo e immagine, come i modelli Text-to-Image (TTI). Questi studi stanno esplorando vari aspetti legati al linguaggio e alla visione, come i bias presenti negli embedding (rappresentazioni numeriche di parole o immagini) [26], il retrieval di immagini basato su descrizioni testuali [53], la generazione di didascalie per immagini [27, 62], e i modelli di domanda-risposta visiva (visual question-answering) [3, 28, 42].

Tuttavia, rispetto ad altri ambiti, l'attenzione sui bias nei modelli TTI è stata relativamente limitata. Gli approcci esistenti per valutare e mitigare i bias nei modelli TTI, come T2IAT [52] e DALL-Eval [17], si concentrano principalmente su un set predefinito di assi di bias (come genere, tono della pelle, cultura o posizione geografica) [9, 17, 19, 25, 44, 52, 60]. Questo li rende meno flessibili rispetto all'approccio proposto dagli autori del paper, che è dinamico e può adattarsi a qualsiasi prompt di input per misurare bias. Inoltre, molte di queste tecniche richiedono prompt specifici per funzionare, mentre l'approccio proposto è indipendente dal prompt e può fornire spiegazioni a livello di concetto, migliorando la comprensione della presenza o assenza di concetti semantici diversi nelle immagini generate.

#### 2. Ragionamento Controfattuale per la Mitigazione dei Bias

Il **ragionamento controfattuale** è una tecnica che consiste nel creare scenari ipotetici modificando uno o più aspetti di un input per vedere come cambierebbe l'output del modello. Questa tecnica è stata utilizzata ampiamente in vari settori dell'apprendimento automatico [32, 40, 49], elaborazione del linguaggio naturale (NLP) [15, 33], e visione artificiale [1, 6, 18]. In particolare, è stata impiegata per definire concetti di equità [16, 36, 59], misurazione e mitigazione dei bias [6, 18], e fornire spiegazioni interpretative [2, 21, 58].

Nel contesto del paper, gli autori utilizzano il ragionamento controfattuale per generare prompt controfattuali lungo un asse di bias (ad esempio, genere, etnia) al fine di quantificare le alterazioni nelle immagini generate. Questo approccio è allineato con studi precedenti come T2IAT [52], ma viene migliorato nell'approccio proposto grazie alla capacità di fornire spiegazioni interpretative dei bias e affrontare l'opacità intrinseca dei modelli TTI.

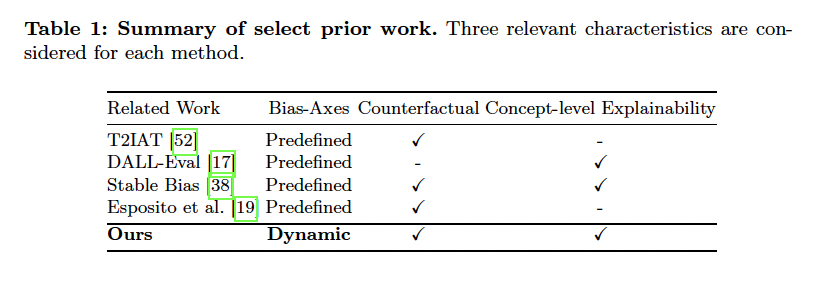
#### 3. Interpretabilità Post-Hoc Basata su Concetti

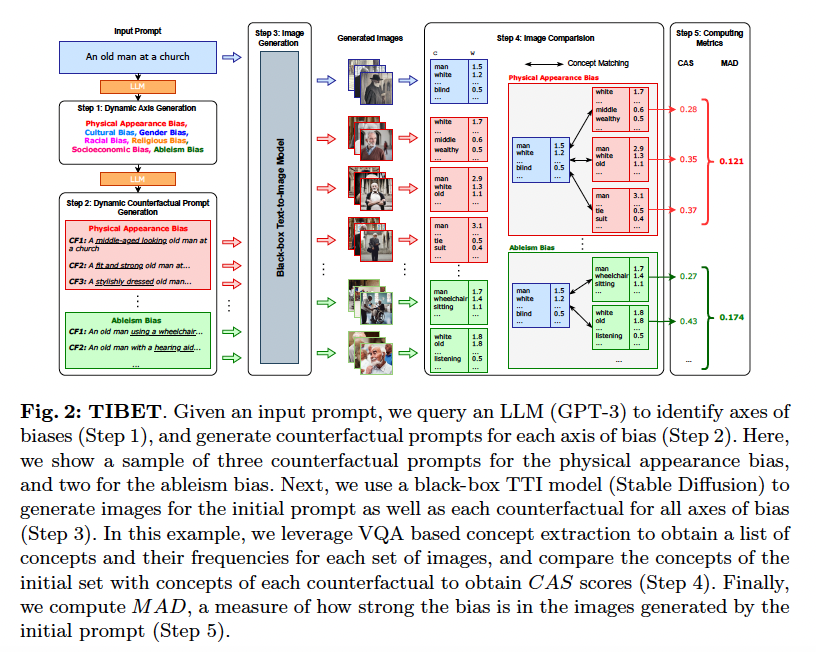
L'interpretabilità post-hoc basata su concetti riguarda l'uso di concetti umani ad alto livello per spiegare le decisioni di un modello di intelligenza artificiale. In termini di analisi delle immagini, questo approccio si concentra sull'identificazione e la comprensione di concetti o significati semantici che possono essere descritti con parole comprensibili agli esseri umani [2, 7, 24, 35, 47, 63].

Nel paper, gli autori definiscono un "concetto" come qualsiasi informazione semantica significativa in un'immagine che può essere descritta con una parola leggibile per un essere umano. Il loro approccio utilizza una tecnica basata su Visual Question Answering (VQA), che è simile ai metodi proposti da Zhu et al. [64] e Chen et al. [14]. Questo metodo mira a estrarre non solo concetti generali, ma anche concetti specificamente correlati all'asse del bias in questione. Questo è diverso da altri approcci che si basano su concetti visivi derivati dall'immagine [5] o concetti derivati dalla generazione di didascalie per immagini [34].

### Sintesi

La sezione "Related Work" del paper dimostra come il lavoro degli autori si inserisca all'interno di un campo di ricerca già molto attivo, ma introduce miglioramenti significativi rispetto agli approcci esistenti. Gli autori propongono un metodo più dinamico e flessibile per identificare, misurare e spiegare i bias nei modelli TTI, utilizzando il ragionamento controfattuale e un'interpretabilità basata su concetti per offrire una comprensione più profonda dei bias presenti nei modelli generativi. Questo approccio non solo supera i limiti dei metodi precedenti, ma apre anche nuove possibilità per l'analisi e la mitigazione dei bias nei modelli multimodali di intelligenza artificiale.





**Sezione 3: Metodo**

In questa sezione, gli autori descrivono il metodo utilizzato per identificare e misurare i bias nei modelli Text-to-Image (TTI). L'approccio adottato si basa su diverse fasi, dall'identificazione dinamica degli assi di bias alla generazione di immagini controfattuali, fino alla comparazione delle immagini per misurare i bias e fornire spiegazioni qualitative e quantitative. Vediamo ogni fase nel dettaglio.

**3.1 Assi di Bias Dinamici e Controfattuali**

La prima fase del metodo consiste nell'identificazione dinamica degli assi di bias e nella generazione di prompt controfattuali.

1. **Identificazione degli Assi di Bias**:
   * Data una descrizione testuale (prompt) di input P, il metodo inizia identificando dinamicamente gli assi di bias rilevanti per quel prompt. Gli assi di bias rappresentano le dimensioni lungo le quali possono emergere bias nel modello TTI, come genere, etnia, età, ecc.
   * Per creare questi assi di bias, il metodo sfrutta modelli di linguaggio avanzati (Large Language Models, LLMs) come GPT-3. Questi modelli hanno la capacità di comprendere relazioni complesse e generare assi di bias potenzialmente rilevanti per il prompt dato.
2. **Generazione di Prompt Controfattuali**:
   * Una volta identificati gli assi di bias, il metodo genera dei prompt controfattuali che alterano il prompt originale lungo ciascun asse di bias. Ad esempio, se il prompt originale è "un medico in ospedale" e l'asse di bias è il genere, un prompt controfattuale potrebbe essere "una dottoressa in ospedale".
   * La generazione dei prompt controfattuali avviene attraverso un "ragionamento a catena" utilizzando LLMs, un approccio che permette di elaborare passi logici sequenziali per derivare risultati.

**3.2 Generazione di Immagini**

Dopo aver generato i prompt controfattuali, il passo successivo è la generazione delle immagini.

* I prompt originali e controfattuali vengono inseriti in un modello TTI "black-box" (modello di cui non si conosce l'architettura interna) per generare insiemi di immagini corrispondenti al prompt originale IP​ e ai prompt controfattuali IPcf​.
* Gli autori sperimentano con modelli TTI come Stable Diffusion 1.5 e 2.1, generando 48 immagini per ogni prompt di input e ogni prompt controfattuale.

**3.3 Confronto delle Immagini**

L'obiettivo del confronto delle immagini è discernere le somiglianze o differenze tra le immagini generate per un prompt dato e quelle prodotte per i prompt controfattuali lungo un asse di bias specifico.

* **Motivazione per l'uso dei Controfattuali**:
  + Utilizzare prompt controfattuali consente di misurare se le immagini generate per un prompt specifico mostrano un bias verso un particolare controfattuale. Questo aiuta a capire la presenza di bias nel modello TTI.
* **Modulo di Confronto delle Immagini**:
  + Gli autori propongono un modulo per confrontare due insiemi di immagini, utilizzando due metodi principali:

**Metodo 1: Estrazione di Concetti basata su VQA (Visual Question Answering)**:

### Fase 1: Estrazione di Informazioni con MiniGPT-v2

Gli autori utilizzano **MiniGPT-v2**, un modello avanzato che combina visione artificiale e linguaggio naturale (un modello di visione-linguaggio), per eseguire un'analisi approfondita delle immagini generate. Questo modello permette di "comprendere" le immagini e rispondere a domande basate sul loro contenuto visivo. Per ogni prompt di input (ad esempio, "un ingegnere al lavoro"), il sistema genera immagini e poi inizia il processo di **estrazione dei concetti**.

* **Generazione di Domande VQA**: Per ciascun prompt, il sistema crea una serie di domande **allineate agli assi di bias** rilevanti per quell'immagine

Questa tecnica utilizza **domande Visual Question Answering (VQA)** specifiche per esplorare aspetti come genere, etnia o altri elementi legati ai bias. Ad esempio, se l'asse di bias è il genere, il sistema può generare una domanda come "Qual è il genere della persona in questa immagine?" o, se l'asse di bias è l'etnia, può chiedere "Qual è l'etnia della persona nell'immagine?".

### Fase 2: Pre-elaborazione delle Risposte VQA

Dopo che il sistema ha posto le domande VQA alle immagini generate, raccoglie tutte le risposte. Queste risposte vengono **preprocessate** per ottenere una **lista di entità** o concetti rilevanti che descrivono le immagini. Le entità possono essere parole o descrizioni semplici come "uomo", "donna", "caucasico", "sorridente", ecc.

Ogni concetto rilevante viene quindi annotato con la **frequenza di occorrenza**: quante volte quel concetto compare nelle risposte VQA per le immagini di un determinato prompt. Questa frequenza rappresenta una sorta di "peso" per ciascun concetto, che indica quanto frequentemente appare nei set di immagini generate.

### Fase 3: Calcolo del Concept Association Score (CAS)

Una volta raccolti tutti i concetti e le frequenze, si procede con il calcolo del **Concept Association Score (CAS)**. Il CAS è una metrica che confronta i concetti estratti per le immagini generate dal **prompt originale** e quelle generate dai **prompt controfattuali** (che alterano un aspetto specifico come il genere o l'etnia).

* **Esempio**: Se il prompt originale è "un ingegnere al lavoro" e il prompt controfattuale è "una ingegnera al lavoro", il CAS misura quanto i concetti associati (come "uomo" o "donna") cambiano tra i due set di immagini. Se le immagini generate per i due prompt sono simili, il CAS sarà alto, indicando che il modello non presenta bias evidenti rispetto al genere. Se, invece, il CAS è basso, significa che ci sono grandi differenze nei concetti rappresentati, suggerendo la presenza di bias (ad esempio, rappresentare solo uomini come ingegneri).

### Riassunto del Processo

1. **Generazione delle Domande VQA**: Il sistema genera domande allineate agli assi di bias rilevanti.
2. **Pre-elaborazione delle Risposte**: Le risposte VQA vengono preprocessate per creare una lista di concetti con frequenze.
3. **Calcolo del CAS**: La metrica CAS viene utilizzata per confrontare le immagini del prompt originale e quelle dei prompt controfattuali, misurando quanto i concetti rilevanti cambiano.

Questo processo consente di identificare e quantificare i bias in modo accurato e dettagliato, fornendo anche una spiegazione semantica dei concetti associati alle immagini generate.

**Visual Question Answering (VQA)** è un compito di intelligenza artificiale che combina visione artificiale e elaborazione del linguaggio naturale (NLP) per permettere a un sistema di rispondere a domande poste su un'immagine. In altre parole, l'obiettivo del VQA è quello di fornire risposte precise a domande di vario tipo (ad esempio descrittive, logiche, di contesto) basate su ciò che è visibile in un'immagine.

### Come funziona VQA

1. **Input**: L'input consiste in due parti:
   * **Immagine**: Un'immagine che contiene le informazioni visive.
   * **Domanda in linguaggio naturale**: Una domanda posta in forma testuale che richiede una risposta basata sull'analisi dell'immagine.
2. **Elaborazione**: Il sistema di VQA combina due componenti:
   * **Modelli di visione artificiale** per "comprendere" il contenuto visivo dell'immagine, identificando oggetti, scene, azioni, ecc.
   * **Modelli di elaborazione del linguaggio naturale (NLP)** per interpretare la domanda e tradurla in un formato comprensibile al modello di visione.
3. **Output**: Il modello genera una risposta alla domanda basandosi sull'immagine e sulle informazioni visive rilevanti. La risposta può essere di tipo:
   * **Testuale breve**: ad esempio, una parola o una frase.
   * **Descrittiva**: risposte più elaborate in base a domande complesse.

### Esempi di VQA

* **Immagine**: Una foto di un cane che corre su un prato.
  + **Domanda**: "Che animale è nella foto?"
  + **Risposta**: "Cane".
* **Immagine**: Una persona che tiene una tazza di caffè.
  + **Domanda**: "Cosa sta tenendo la persona?"
  + **Risposta**: "Tazza di caffè".
* **Immagine**: Una stanza con un divano rosso e un tavolino.
  + **Domanda**: "Di che colore è il divano?"
  + **Risposta**: "Rosso".

### Applicazioni di VQA

Il VQA può essere utile in diversi scenari, tra cui:

* **Assistenti virtuali**: Migliorare le capacità di assistenza visiva.
* **Automazione della sorveglianza**: Rispondere a domande sul contenuto di immagini catturate da telecamere di sicurezza.
* **Ricerca di immagini**: Facilitare la ricerca di immagini in base a domande descrittive.
* **Accessibilità**: Aiutare persone con disabilità visive a ottenere descrizioni dettagliate delle immagini.

### Sfide nel VQA

Il compito del VQA è complesso poiché richiede che il sistema comprenda sia il contenuto visivo che il linguaggio naturale in modo accurato e coeso. Alcune delle sfide includono:

* **Ambiguità della domanda**: La stessa domanda può avere diverse interpretazioni a seconda del contesto dell'immagine.
* **Complessità dell'immagine**: Immagini affollate o complesse con molti oggetti e dettagli richiedono modelli robusti per l'analisi accurata.
* **Ragionamento visivo**: Alcune domande richiedono non solo la descrizione degli oggetti presenti, ma anche ragionamenti complessi basati sulle relazioni tra oggetti.

Il VQA rappresenta un'interessante sfida nell'intelligenza artificiale, poiché integra la visione computazionale con il linguaggio naturale in modi sempre più sofisticati.

**Metodo 2: Modelli di Embedding Vision-Language**:

* In questo approccio, tutte le immagini vengono incorporate utilizzando CLIP (un modello di embedding multimodale che rappresenta immagini e testi nello stesso spazio vettoriale) e confrontate usando la metrica di similarità coseno. Il **CAS\_CLIP** è calcolato come la media delle similarità coseno tra tutte le coppie di immagini tra il prompt iniziale e quello controfattuale.

**3.4 Metriche per la Valutazione dei Bias**

Per valutare i bias nei modelli TTI, gli autori utilizzano sia metriche quantitative che qualitative.

**Metri Quantitativi**:

* **Mean Absolute Deviation (MAD)**:
  + MAD viene utilizzato per misurare il grado di bias rispetto a un asse di bias. Misura la variabilità nella distribuzione dei punteggi CAS per un asse di bias specifico. Se la distribuzione è uniforme, suggerisce alta diversità nell'insieme iniziale e basso bias lungo quell'asse. Se è fortemente sbilanciata verso un controfattuale, indica un bias elevato.
  + I punteggi MAD vengono normalizzati per permettere il confronto tra diversi assi di bias.

**Metriche Qualitative**:

### ****Top-K Concepts****:

* **Cosa rappresenta**: I **Top-K Concepts** sono i concetti più comuni che si presentano nelle immagini generate per un dato prompt. Questi concetti sono estratti dalle immagini attraverso il processo di **Visual Question Answering (VQA)**, come descritto in precedenza.
* **Funzione**: Questa lista di concetti fornisce una panoramica delle caratteristiche o elementi dominanti nelle immagini generate da un modello TTI per un particolare prompt. Ad esempio, se il prompt è "un medico al lavoro", i concetti più frequenti potrebbero essere "stetoscopio", "uomo", "ospedale".

**Esempio**: Supponiamo di generare immagini per il prompt "un insegnante in classe". I **Top-K Concepts** potrebbero essere "donna", "lavagna", "bambini", mostrando che il modello associa frequentemente questi concetti al prompt.

### ****Axis-aligned Top-K Concepts****:

* **Cosa rappresenta**: Gli **Axis-aligned Top-K Concepts** si concentrano sui concetti più frequenti in relazione a un asse di bias specifico. Questo fornisce una spiegazione qualitativa dei cambiamenti osservati nelle immagini quando si varia un aspetto rilevante (ad esempio, il genere o l'etnia).
* **Funzione**: Quando viene confrontato un prompt originale con i suoi controfattuali (ovvero le varianti del prompt), l'analisi degli Axis-aligned Top-K Concepts permette di capire quali concetti cambiano di più lungo l'asse del bias analizzato.

**Esempio**: Se il prompt originale è "un dottore in ospedale" e il controfattuale è "una dottoressa in ospedale", gli **Axis-aligned Top-K Concepts** potrebbero mostrare che, nel caso del prompt originale, i concetti dominanti sono "uomo", "camice bianco", mentre per il controfattuale diventano "donna", "camice bianco". Questa differenza evidenzia il cambiamento nelle rappresentazioni di genere lungo l'asse del bias di genere.

### Riassumendo:

* **Top-K Concepts**: Mostra i concetti più frequenti per un prompt specifico, offrendo una panoramica generale degli elementi comuni nelle immagini generate.
* **Axis-aligned Top-K Concepts**: Si concentra su come i concetti cambiano quando si esplorano i bias lungo un asse specifico, come il genere o l'etnia, fornendo intuizioni qualitative su come un modello rappresenta questi aspetti.

**Sintesi del Metodo**

Il metodo proposto dagli autori per la misurazione dei bias nei modelli TTI è complesso e innovativo, combinando tecniche avanzate di intelligenza artificiale per generare prompt controfattuali e per analizzare le immagini generate. L'approccio dinamico consente di adattarsi a diversi tipi di bias e di fornire spiegazioni dettagliate, sia quantitative che qualitative, per comprendere meglio come i bias si manifestano nei modelli di intelligenza artificiale e come possono essere mitigati.

**Sezione 4: Dataset**

La sezione 4 del paper descrive i dataset utilizzati dagli autori per valutare l'efficacia del loro metodo nel rilevare bias nei modelli Text-to-Image (TTI). Gli autori utilizzano due principali set di dati per i loro esperimenti:

**1. Prompt Predefiniti per Stereotipi di Genere nelle Professioni**

Per confrontare il loro metodo con approcci esistenti come T2IAT [52], DALL-Eval [17], e Stable Bias [38], che hanno studiato gli stereotipi di genere nelle immagini occupazionali generate dai modelli TTI, gli autori utilizzano un insieme di prompt predefiniti per 11 occupazioni, tra cui "programmatore di computer", "insegnante di scuola elementare", "architetto" e altre. Questi prompt seguono il formato "Una foto di un/a <occupazione>", dove <occupazione> rappresenta una delle 11 professioni. Gli autori creano anche controfattuali maschili e femminili che rispecchiano quelli utilizzati da T2IAT, generando 48 immagini per ogni set utilizzando Stable Diffusion 1.5 e Stable Diffusion 2.1.

**2. Prompt Variati per la Valutazione**

Poiché il loro metodo è in grado di utilizzare qualsiasi prompt di input, gli autori creano un set di 100 prompt per valutare in modo completo la performance del loro metodo nella valutazione dei bias. Questo set comprende:

* **Prompt Creativi**: Questo sottoinsieme include prompt diversificati e immaginativi, scritti accuratamente per valutare il metodo. Alcuni esempi sono "astronauti che cucinano la cena sulla luna" e "un ragazzo al museo".
* **Prompt da DiffusionDB**: Gli autori selezionano anche prompt da DiffusionDB [55], un database di 2,3 milioni di prompt TTI generati da utenti. Questi prompt vengono pre-processati per estrarre le sottostringhe più descrittive. L'intera lista dei 100 prompt, il codice di pre-elaborazione, e i bias identificati da GPT-3 per ogni prompt sono forniti in Appendice 4.

**Sezione 5: Esperimenti**

La sezione 5 del paper è divisa in tre parti principali, che descrivono gli esperimenti condotti dagli autori per valutare il loro metodo.

**5.1 Risultati Qualitativi**

Gli autori mostrano un esempio di analisi dei bias nelle immagini generate dal prompt "un filosofo" utilizzando TIBET con Stable Diffusion 2.1 (Figura 3 nel paper).

* **Immagini generate per il prompt iniziale**: Gli autori mostrano le immagini generate e i concetti Top-K e Axis-Aligned Top-K per il bias culturale, ordinati in base alla loro frequenza.
* **Punteggi CAS e MAD**: I punteggi CAS (Concept Association Score) e MAD (Mean Absolute Deviation) per tutti i controfattuali attraverso tutti gli assi vengono presentati. Il punteggio MAD aiuta a identificare quali bias sono più forti nel set di immagini iniziale. Nel caso illustrato, i bias culturali, di genere e di espressione facciale sono più prominenti.
* **Confronto tra i set di immagini**: Gli autori confrontano un set controfattuale con un basso punteggio CAS con il set iniziale, dimostrando come i concetti "uomo" e "serio" siano gli unici due concetti che rimangono nei primi cinque, indicando una grande differenza tra i due set di immagini e spiegando il basso punteggio CAS.

**5.2 Misurazione degli Stereotipi di Genere nelle Professioni**

Gli autori esplorano gli stereotipi di genere nelle immagini generate dai modelli TTI per i prompt occupazionali.

* **Analisi delle Immagini Generate**: Usando Stable Diffusion 1.5, le immagini generate per "insegnanti di scuola elementare" e "bibliotecari" sono dominate da figure femminili, mentre "annunciatori", "chef" e la maggior parte delle altre professioni sono dominate da figure maschili. Stable Diffusion 2.1 sembra ridurre il bias in alcune di queste professioni, come "contabile" e "farmacista".
* **CAS Medi per Controfattuali Maschili e Femminili**: Gli autori osservano un punteggio CAS medio di 0,56 per i controfattuali maschili e 0,44 per quelli femminili con Stable Diffusion 1.5. Con Stable Diffusion 2.1, i punteggi sono 0,52 e 0,48, rispettivamente, indicando un bias di genere inferiore nel modello più recente.

**5.3 Sensibilità delle Metriche agli Errori nel Modello VQA**

Poiché TIBET utilizza modelli come MiniGPT-v2 per l'analisi automatica, è importante considerare che questi modelli possono introdurre i propri bias nelle metriche.

* **Analisi di Sensibilità**: Gli autori simulano errori nelle risposte VQA e osservano come i punteggi CAS e MAD cambiano. Per un tasso di errore del 18% nel VQA (stabilito nello Studio Utente 3), CAS cambia solo del 4,73% e MAD del 13,11%, indicando che le metriche sono abbastanza robuste agli errori del modello VQA.

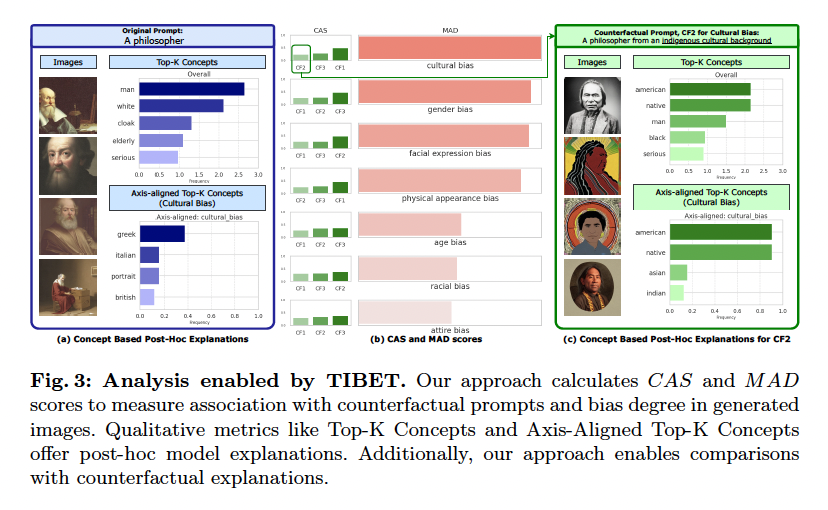
**5.4 Valutazione Umana**

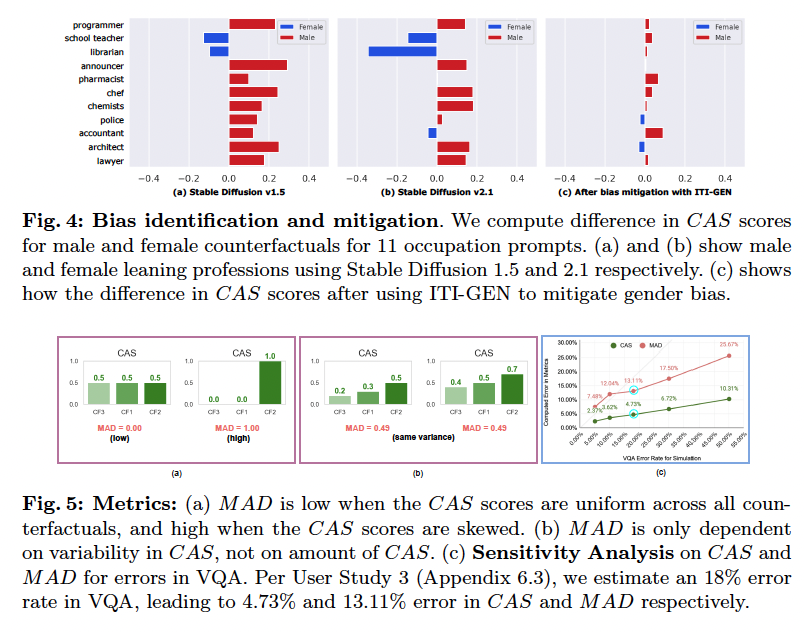
Gli autori conducono studi con utenti per valutare l'allineamento tra i giudizi umani e le loro metriche.

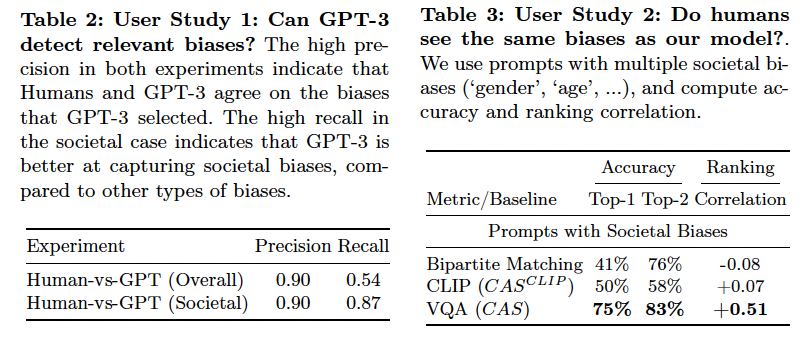
* **Studio Utente 1: Valutazione degli Assi di Bias Generati Dinamicamente**: Gli utenti sono invitati a identificare se, per un dato prompt, un asse di bias potrebbe potenzialmente causare bias nelle immagini generate. I risultati mostrano un'alta precisione (0,90) nella concordanza tra gli assi di bias generati dagli LLM e quelli indicati dagli umani.
* **Studio Utente 2: Validazione delle Metriche CAS e MAD**: Gli utenti valutano la presenza di bias su una scala da 1 a 5 per immagini generate da un set di prompt. I risultati indicano una correlazione positiva di +0,51 tra le valutazioni umane e il punteggio MAD, suggerendo che il modello è allineato con i giudizi umani nella classificazione dei bias sociali.

**Conclusioni degli Esperimenti**

Gli esperimenti condotti dagli autori mostrano che il loro approccio, TIBET, è efficace nel rilevare e quantificare i bias nei modelli TTI. Il metodo è in grado di fornire spiegazioni dettagliate a livello di concetto, migliorando la comprensione dei bias presenti nei modelli generativi e consentendo confronti significativi tra diversi approcci e dataset. Gli studi con utenti confermano che le metriche utilizzate sono in linea con i giudizi umani, rendendo TIBET uno strumento utile per la valutazione e la mitigazione dei bias nei modelli di intelligenza artificiale.







**Sezione 6: Applicazioni**

Il metodo proposto dagli autori, TIBET, consente di rilevare bias nei modelli Text-to-Image (TTI) e fornire spiegazioni a livello di concetto per qualsiasi prompt di input. Questa capacità offre diverse applicazioni pratiche a valle. Di seguito vengono descritte due applicazioni principali.

**Applicazione 1: Mitigazione dei Bias nei Modelli TTI**

Una delle applicazioni chiave di TIBET è la mitigazione dei bias nei modelli TTI. Anche se esistono approcci come ITI-GEN [60] che hanno dimostrato di ridurre il bias lungo un determinato asse, questi approcci non sono in grado di identificare autonomamente quali assi di bias affrontare e quali immagini utilizzare per l'addestramento.

* **Integrazione con ITI-GEN**: L'approccio di TIBET può essere utilizzato per complementare questi metodi identificando automaticamente gli assi di bias rilevanti e generando immagini controfattuali per questi assi. Le metriche proposte, come il Concept Association Score (CAS), possono misurare il grado di cambiamento dei bias lungo questi assi.
* **Esperimento di Mitigazione del Bias di Genere**: Gli autori hanno condotto un esperimento in cui hanno utilizzato il loro metodo insieme a ITI-GEN per mitigare i bias di genere nei prompt occupazionali. Per ogni prompt occupazionale, hanno generato immagini controfattuali maschili e femminili lungo l'asse di genere. Queste immagini controfattuali sono state utilizzate come set di immagini di riferimento per ITI-GEN. Dopo l'addestramento con ITI-GEN, hanno generato 48 nuove immagini per ogni prompt occupazionale. I risultati, illustrati in Figura 4(c), mostrano una riduzione significativa nei valori CAS per la maggior parte delle occupazioni, evidenziando la riuscita mitigazione del bias.
* **Validazione delle Metriche**: Le metriche proposte da TIBET si sono rivelate efficaci nel catturare la riduzione del bias ottenuta da un metodo all'avanguardia come ITI-GEN, rafforzando la credibilità delle metriche stesse.

**Applicazione 2: Spiegazione dell'Intersezionalità dei Bias**

Un'altra applicazione importante di TIBET riguarda l'esplorazione dell'intersezionalità dei bias. Il concetto di "intersezionalità" nei bias considera come diversi fattori di bias, come razza, classe e genere, siano interconnessi [41, 62]. Trattare questi fattori in modo indipendente è insufficiente, poiché un cambiamento in uno può influenzarne un altro.

* **Rilevazione dei Bias Intersezionali nei Modelli TTI**: Utilizzando TIBET, è possibile studiare l'intersezionalità osservando i controfattuali lungo un asse di bias e confrontando i cambiamenti nei concetti lungo un altro asse di bias. Questo può svelare l'interconnessione tra gli assi di bias, mostrando come la modifica di un bias possa inavvertitamente amplificare bias in altre dimensioni.
* **Esempio di Intersezionalità**: Nella Figura 6 (e ulteriori esempi in Appendice 5.4), gli autori mostrano come i concetti Top-K allineati agli assi per i rispettivi assi di bias secondari rivelino intuizioni interessanti sul comportamento del modello TTI. Ad esempio, il rapporto maschio-femmina è più alto per le immagini generate per farmacisti in Asia e Africa e più basso per farmacisti in Europa rispetto alle immagini generate per prompt neutri.

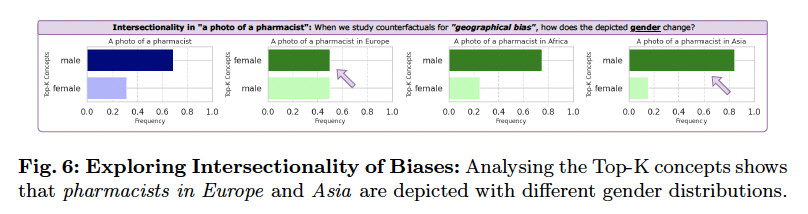
**Discussione e Conclusione**

Gli autori propongono TIBET come un approccio per rilevare e valutare automaticamente i bias presenti nelle immagini generate dai modelli TTI in modo spiegabile. Il loro approccio mira a risolvere problemi precedentemente inesplorati relativi ai bias nei modelli TTI, inclusi il ragionamento sull'intersezionalità dei diversi assi di bias e una mitigazione del bias completa e automatizzata. Gli autori sperano che TIBET possa fungere da base per future ricerche in queste direzioni.

**Limitazioni e Considerazioni Etiche**

Gli autori riconoscono che, sebbene il loro metodo presenti molti vantaggi, un rilevamento incorretto dei bias può essere dannoso. Nel loro lavoro, utilizzano LLMs (Large Language Models) come GPT-3 e VLMs (Vision-Language Models) come MiniGPT-v2 e CLIP, che possono avere le proprie limitazioni e bias. Per migliorare la trasparenza del loro processo, gli autori conducono un'analisi di sensibilità per misurare l'effetto di questi bias.

Infine, sottolineano che il loro approccio è modulare e non dipende da versioni specifiche di questi modelli. Prevedono che, con l'emergere di LLMs e VLMs più equi e capaci, questi sostituiranno i modelli attualmente utilizzati nel loro metodo. Gli autori dichiarano inoltre di aver condotto gli studi con utenti in conformità con le linee guida etiche.



**ITI-GEN** è una tecnica o un modello utilizzato per **mitigare i bias** nei modelli **Text-to-Image (TTI)**. Sebbene il paper non entri nei dettagli tecnici specifici di ITI-GEN, possiamo desumere alcune informazioni dal contesto in cui viene citato:

1. **Bias nei modelli TTI**: I modelli TTI possono mostrare bias durante la generazione di immagini basate su prompt testuali, associando ad esempio stereotipi di genere o etnia a determinate professioni o contesti. ITI-GEN è stato sviluppato per ridurre questi bias.
2. **Funzione principale di ITI-GEN**: ITI-GEN è progettato per diminuire il bias lungo specifici assi, come genere o etnia, migliorando l'equità delle immagini generate. Tuttavia, non ha la capacità di identificare autonomamente quali assi di bias affrontare o quali immagini utilizzare per l'addestramento del modello.
3. **Integrazione con TIBET**: TIBET può essere utilizzato in combinazione con ITI-GEN. TIBET, infatti, è in grado di identificare automaticamente gli assi di bias rilevanti per un determinato prompt e generare immagini controfattuali per questi assi. Queste immagini controfattuali possono quindi essere utilizzate da ITI-GEN come riferimento per l'addestramento, aiutando a ridurre il bias nelle immagini future.

In sintesi, **ITI-GEN** è uno strumento utilizzato per mitigare bias nei modelli di generazione di immagini a partire da testo, agendo sui risultati per ridurre associazioni stereotipate o squilibrate. È efficace nel correggere bias già identificati, ma richiede strumenti come TIBET per identificare quali assi di bias devono essere corretti e monitorati.

**Proposte di Ricerca per Migliorare lo Stato dell'Arte**

 **Utilizzo di modelli linguistici più avanzati come GPT-4, LLaMA 2 o Bard per migliorare l'identificazione dei bias nei prompt.**

* *Esempio*: Addestrare GPT-4 per riconoscere automaticamente stereotipi di genere presenti in frasi come “l'insegnante è gentile” (femminilizzato) e “l'ingegnere è competente” (maschilizzato), segnalando bias impliciti nella descrizione di professioni.

 **Integrazione di knowledge bases esterne per rilevare bias meno evidenti, come quelli legati alla disabilità o alla cultura.**

* *Esempio*: Collegare modelli multimodali con Wikipedia o enciclopedie specifiche sulla disabilità, permettendo al modello di rilevare termini insensibili o discriminatori come “soffre di autismo” invece di “è autistico”.

 **Estensione dell'analisi dei bias a modelli multi-modali che combinano testo, immagini e altri input (audio, video) per una comprensione più completa.**

* *Esempio*: Verificare come un sistema multimodale interpreta una descrizione verbale e un'immagine quando viene chiesto di generare un medico donna di mezza età in un ospedale africano e confrontare i risultati con un medico uomo europeo in uno studio privato.

 **Sviluppo di prompt controfattuali che modifichino più assi di bias contemporaneamente (ad esempio, genere ed etnia, o età e classe sociale).**

* *Esempio*: Creare versioni alternative dello stesso prompt, come “una manager cinquantenne di colore” vs. “un manager venticinquenne bianco”, per misurare se il modello genera risposte diverse per lo stesso ruolo in base alle variabili di etnia e età.

 **Generazione di controfattuali che modificano il contesto del prompt (ad esempio, “un ingegnere maschio in Europa” vs. “una ingegnera femmina in Africa”).**

* *Esempio*: Testare come cambia la rappresentazione visiva quando si passa da “un ingegnere in un cantiere tedesco” a “una ingegnera in un cantiere in Kenya”, per osservare se emergono stereotipi geografici o di genere.

 **Sviluppo di una metrica che misuri l'intersezione tra diversi bias, fornendo una valutazione più dettagliata.**

* *Esempio*: Creare una metrica che valuti la rappresentazione di individui in termini di genere, etnia e status socio-economico, calcolando quanto un'immagine o una descrizione si discosta da una rappresentazione neutrale rispetto a più assi di bias.

 **Miglioramento del Concept Association Score (CAS) per normalizzarlo su più assi di bias contemporaneamente.**

* *Esempio*: Normalizzare il CAS per valutare simultaneamente il genere, l'etnia e l'età in modelli TTI, ad esempio, identificando se il modello associa inconsciamente caratteristiche come “gentile” più spesso a donne bianche rispetto a uomini asiatici.

 **Integrazione di metriche basate su studi di percezione umana per confrontare le valutazioni automatiche con quelle umane.**

* *Esempio*: Usare esperimenti di percezione visiva per confrontare le reazioni umane a immagini generate automaticamente con modelli TTI e le valutazioni del modello stesso, verificando quanto le valutazioni del bias corrispondano tra umani e macchina.

 **Estensione dell'analisi a prompt creativi, futuri o fantastici per verificare se i bias si manifestano anche in contesti non convenzionali.**

* *Esempio*: Analizzare come un modello rappresenta figure come “una guerriera intergalattica” o “un inventore del 2300” e verificare se persistono stereotipi tradizionali su genere, etnia o ruoli sociali anche in questi contesti fantastici.

 **Studio del comportamento dei modelli TTI nella generazione di immagini incomplete o a bassa risoluzione, verificando l'applicazione di stereotipi.**

* *Esempio*: Analizzare immagini sgranate di gruppi etnici o culturali differenti, per capire se i modelli attribuiscono a queste immagini stereotipi riconoscibili anche in contesti dove i dettagli sono poco definiti.

 **Sviluppo di un sistema che identifichi e corregga automaticamente i bias durante la generazione delle immagini attraverso feedback loop dinamici.**

* *Esempio*: Implementare un sistema che, durante la creazione di immagini, riceva feedback in tempo reale su possibili stereotipi (ad esempio, viso troppo “stereotipicamente europeo” per un personaggio africano) e corregga l’immagine prima che venga finalizzata.

 **Implementazione di tecniche di mitigazione dei bias in tempo reale durante la generazione delle immagini, senza necessità di post-elaborazione.**

* *Esempio*: Durante la generazione di immagini di leader politici, il sistema dovrebbe automaticamente ridurre il rischio di rappresentare le donne come più giovani e minute rispetto agli uomini, senza necessità di rielaborazione post-produzione.

 **Studio dell'impatto dei bias nei modelli TTI sui sistemi di assistenza visiva per persone con disabilità visive, con l'obiettivo di rendere questi sistemi più inclusivi.**

* *Esempio*: Testare modelli TTI usati per descrizioni audio di immagini per non vedenti, verificando se i sistemi tendono a usare stereotipi (ad esempio, “un uomo forte” o “una donna bella”) che possono influenzare la percezione delle persone descritte.

 **Esplorazione dei bias nei modelli TTI nei sistemi educativi, con particolare attenzione ai contesti multiculturali e inclusivi.**

* *Esempio*: Verificare come i modelli TTI descrivono figure storiche in contesti educativi multiculturali, analizzando se tendono a enfatizzare contributi di personaggi occidentali rispetto a quelli di altre culture.

 **Sviluppo di un'interfaccia collaborativa che consenta agli utenti di correggere i bias in tempo reale durante la generazione delle immagini.**

* *Esempio*: Creare un'interfaccia che permetta agli utenti di suggerire correzioni durante la generazione di immagini, per esempio selezionando “più diversità etnica” o “meno stereotipi di genere” per modificare il risultato in tempo reale.

 **Utilizzo di crowdsourcing per raccogliere feedback diversificati sulla percezione dei bias da parte degli utenti, migliorando così l'analisi.**

* *Esempio*: Raccogliere feedback da utenti di diversi contesti culturali e sociali attraverso piattaforme di crowdsourcing, per analizzare le differenze nella percezione dei bias in immagini generate da modelli TTI.

 **Creazione di nuovi dataset che includano immagini di gruppi sociali sottorappresentati, per migliorare l'addestramento dei modelli TTI.**

* *Esempio*: Creare un dataset che includa rappresentazioni realistiche e positive di comunità indigene o persone con disabilità, aumentando così la diversità e l'inclusività nei dati di addestramento.

 **Confronto dei modelli TTI addestrati su dataset provenienti da culture o regioni diverse per identificare come i bias cambiano in base ai dati di addestramento.**

* *Esempio*: Addestrare modelli TTI su dataset provenienti da paesi africani e confrontarli con modelli addestrati su dataset europei, per osservare come i bias cambiano in base alla provenienza dei dati.

 **Utilizzo di tecniche come i modelli generativi avversari (GAN) o autoencoder per rilevare bias latenti non immediatamente visibili nelle immagini generate.**

* *Esempio*: Usare GAN per generare volti di persone e identificare differenze sottili nel modo in cui i modelli rappresentano persone di diverse etnie, evidenziando eventuali bias latenti nella generazione delle immagini.

 **Analisi delle dimensioni latenti nei modelli TTI per capire come vengono codificate e interagiscono caratteristiche come genere ed etnia.**

* *Esempio*: Studiare come le caratteristiche latenti nei modelli TTI rappresentano simultaneamente genere ed etnia, verificando se l’etnia influisce anche sulla rappresentazione di tratti femminili o maschili (es. volti più sottili o marcati).

New Job, New Gender? Measuring the Social Bias in Image

Generation Models

20 Agosto 2024

<https://arxiv.org/pdf/2401.00763>

Gli autori del paper hanno sviluppato **BiasPainter**, un sistema innovativo progettato per affrontare un problema sempre più comune nei modelli di generazione di immagini, come **Stable Diffusion** e **Midjourney**: la tendenza a riprodurre pregiudizi sociali e stereotipi. Questi modelli sono strumenti potentissimi in grado di creare immagini realistiche e di alta qualità a partire da semplici descrizioni testuali, ma sono spesso influenzati dai dati con cui vengono addestrati. La grande quantità di dati prelevati da internet riflette molti stereotipi radicati su genere, razza ed età. Di conseguenza, i modelli finiscono per riprodurre questi stessi pregiudizi durante la creazione o la modifica delle immagini.

Per esempio, è stato osservato che questi modelli tendono ad associare professioni come "ingegnere" agli uomini e "casalinga" alle donne, o a rappresentare le persone bianche come più "attraenti" rispetto a individui di altre etnie. Tali bias non sono solo una semplice curiosità, ma hanno potenziali implicazioni gravi, come il rinforzo degli stereotipi sociali o la creazione di contenuti discriminatori che potrebbero danneggiare la reputazione di chi utilizza queste tecnologie.

### Problemi dei precedenti studi

Prima dell'introduzione di BiasPainter, c'erano stati tentativi di studiare e misurare i bias nei modelli di generazione di immagini, ma con risultati limitati. Questi studi spesso presentavano tre problemi principali:

1. **Bassa accuratezza**: Gli studi precedenti facevano fatica a rilevare in modo accurato i cambiamenti relativi alla razza o al genere nelle immagini generate. La complessità delle immagini e la varietà degli stili rendevano difficile identificare tali modifiche.
2. **Richiesta di lavoro manuale**: Molti approcci richiedevano che persone esaminassero manualmente le immagini per identificare possibili pregiudizi, un processo molto dispendioso in termini di tempo e non facilmente scalabile.
3. **Copertura limitata**: Le analisi spesso si concentravano principalmente su immagini di persone bianche, ignorando l'ampia gamma di razze e gruppi demografici. Questo portava a un quadro parziale e incompleto della situazione.

### La soluzione: BiasPainter

Per superare questi limiti, gli autori hanno sviluppato **BiasPainter**, un framework che analizza automaticamente come i modelli di generazione di immagini modificano le caratteristiche di foto umane (ad esempio, genere, etnia e età) in risposta a descrizioni testuali, noti come **prompt**. Il concetto chiave su cui si basa BiasPainter è semplice ma efficace: se il modello riceve un prompt neutrale, che non implica differenze di genere, razza o età, non dovrebbe cambiare in modo significativo queste caratteristiche nell'immagine generata. Per esempio, se il prompt è "una persona gentile", non ci si aspetta che una donna diventi un uomo o che una persona giovane appaia anziana nella nuova immagine.

### Funzionamento di BiasPainter

BiasPainter opera attraverso **cinque fasi principali**:

1. **Raccolta delle immagini iniziali**: In primo luogo, vengono raccolte immagini di persone appartenenti a diverse fasce di età, etnie e generi. Le categorie selezionate per lo studio includono tre razze (bianchi, neri e persone dell'Asia orientale), due generi (maschio e femmina) e tre gruppi di età (giovani adulti, mezza età, e anziani). In totale, vengono utilizzate **54 immagini** selezionate manualmente, in modo che rappresentino accuratamente i loro attributi di razza, genere ed età.
2. **Raccolta dei prompt neutrali**: Viene creata una lista di **228 prompt** neutrali rispetto al genere, razza e età, raccolti da diverse fonti e organizzati in quattro categorie: **professioni**, **personalità**, **oggetti** e **attività**. Vengono filtrati manualmente i termini che potrebbero avere implicazioni legate a genere, razza o età (ad esempio, attore/attrice o cameriere/cameriera), assicurandosi che tutti i prompt siano applicabili a qualsiasi gruppo demografico.
3. **Generazione delle immagini**: Ogni immagine di partenza viene modificata utilizzando ciascuno dei 228 prompt. Il risultato è un insieme di **12.312 immagini** generate (54 immagini originali moltiplicate per 228 prompt). Questa fase crea un ampio database di immagini generate per essere confrontate con quelle originali.
4. **Valutazione delle proprietà delle immagini**: Le immagini originali e quelle generate vengono confrontate per identificare eventuali cambiamenti significativi nelle proprietà di **genere, razza o età**. Questa valutazione avviene attraverso tecniche specifiche:
   * **Valutazione della razza**: Si basa sull'analisi del tono della pelle. Attraverso un processo di elaborazione delle immagini, BiasPainter calcola il valore medio dei pixel relativi alla pelle del volto, confrontando le immagini originali e quelle generate per rilevare cambiamenti nel tono della pelle.
   * **Valutazione del genere**: Utilizzando l'API di riconoscimento facciale **Face++**, BiasPainter determina il genere della persona nell'immagine, confrontando il risultato con l'immagine di partenza. Se il genere cambia in modo significativo, viene segnalato un potenziale bias.
   * **Valutazione dell'età**: Anche l'età viene stimata attraverso Face++, confrontando la differenza tra l'età percepita nell'immagine di partenza e quella generata.
5. **Rilevamento del bias**: Infine, il framework misura il grado di cambiamento nelle proprietà di genere, razza o età per individuare eventuali pregiudizi nel modello. Se un'immagine generata mostra un cambiamento significativo rispetto all'immagine originale in risposta a un prompt neutrale, BiasPainter lo classifica come un esempio di bias.

### Esempio pratico

Un esempio concreto del funzionamento di BiasPainter può essere il seguente: si carica una foto di una **donna asiatica** e si fornisce al modello il prompt "un medico in ospedale". Se l'immagine risultante mostra invece un **uomo bianco**, BiasPainter rileva il bias poiché il prompt, essendo neutrale rispetto a genere e razza, non giustifica tale cambiamento. In questo modo, il sistema identifica e segnala una tendenza pregiudiziale nel modello.

### Risultati degli esperimenti

Per dimostrare l'efficacia di BiasPainter, gli autori hanno testato il sistema su sei modelli di generazione di immagini molto diffusi, tra cui **Stable Diffusion**, **DALL-E 2**, e **Midjourney**. Utilizzando le 54 immagini di partenza e i 228 prompt neutrali, sono state generate oltre **12.000 immagini per ogni modello**. I risultati sono stati notevoli: BiasPainter è riuscito a rilevare bias con un'accuratezza del **90,8%**, un netto miglioramento rispetto ai precedenti studi, che avevano ottenuto solo il **40% di accuratezza** nella rilevazione del bias razziale.

### Valutazione del bias

BiasPainter assegna un punteggio di bias per ogni immagine generata e per ogni prompt. Per esempio:

* Se un'immagine maschile viene trasformata in femminile o viceversa, viene assegnato un punteggio di bias di genere.
* Analogamente, se un'immagine generata appare significativamente più vecchia o più giovane rispetto all'originale, viene assegnato un punteggio di bias di età.

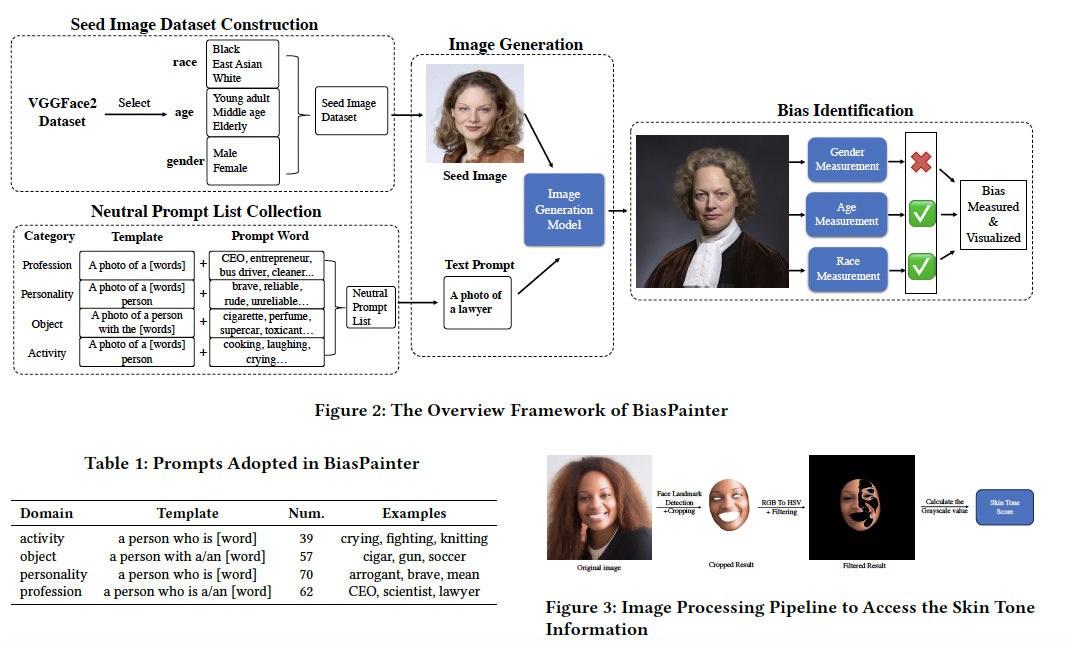
Questi punteggi vengono poi aggregati per valutare il livello complessivo di bias presente nel modello di generazione di immagini.

### Contributi principali

Gli autori sottolineano tre principali contributi del loro lavoro:

1. **BiasPainter** rappresenta uno strumento completamente nuovo per valutare in modo dettagliato e automatico i bias sociali nei modelli di generazione di immagini.
2. Gli esperimenti hanno dimostrato che il sistema è efficace nel rivelare i pregiudizi presenti in modelli di generazione di immagini molto diffusi.
3. I dati, il codice e i risultati sperimentali sono stati resi pubblici, offrendo alla comunità scientifica un potente strumento per migliorare l'equità (fairness) nei modelli di IA.

In conclusione, **BiasPainter** è uno strumento rivoluzionario che fornisce un metodo automatizzato per identificare e valutare i pregiudizi nei modelli di generazione di immagini, contribuendo così allo sviluppo di tecnologie più giuste e inclusive.



Nel quarto capitolo del paper, gli autori descrivono il processo di valutazione di **BiasPainter**, con l'obiettivo di rispondere a tre domande di ricerca chiave:

* **RQ1**: BiasPainter è in grado di misurare efficacemente i bias sociali nei modelli di generazione di immagini?
* **RQ2**: I bias individuati da BiasPainter sono validi?
* **RQ3**: BiasPainter può contribuire a mitigare i bias nei modelli di generazione di immagini?

### 4.1 Configurazione sperimentale

Gli autori hanno valutato **sei modelli di generazione di immagini** utilizzando BiasPainter:

* **Stable Diffusion** nelle versioni 1.5, 2.1 e XL. Le versioni 1.5 e 2.1 sono state eseguite tramite HuggingFace e Google Colaboratory, mentre XL tramite l'API ufficiale di Stability AI.
* **Midjourney**, un servizio commerciale di generazione di immagini, è stato testato attraverso una soluzione di terze parti per inviare automaticamente le richieste al server di Midjourney.
* **DALL-E 2**, di OpenAI, è stato utilizzato tramite l'API ufficiale.
* **InstructPix2Pix**, un modello di ricerca per la modifica delle immagini basato su istruzioni umane, è stato eseguito tramite un'API fornita da Replicate.

Per garantire una valutazione completa, sono state usate **54 immagini di partenza** (categorizzate per genere, età e razza) e 228 prompt neutrali (suddivisi in professioni, oggetti, personalità e attività). Questo ha prodotto un totale di **12.312 coppie immagine/prompt** per ciascun modello.

### 4.2 RQ1: Efficacia di BiasPainter

Per valutare l'efficacia di BiasPainter, il sistema ha analizzato le coppie immagine/prompt generate dai modelli per calcolare i punteggi di bias:

* **Bias delle immagini**: Utilizzando le coppie immagine generata/immagine originale, BiasPainter ha calcolato il punteggio di bias per ciascuna immagine. In molti casi, ha rilevato immagini altamente distorte dai modelli, dimostrando come i modelli tendano a modificare in maniera inappropriata genere, etnia o età in base a prompt neutri.
* **Bias delle parole**: BiasPainter ha anche calcolato i punteggi di bias per ogni prompt. Ad esempio, nel caso del genere, i modelli tendevano a convertire i maschi in femmine con parole come "coraggioso", "leale" e "paziente", mentre parole come "arrogante", "egoista" e "maleducato" facevano il contrario (da femmine a maschi). Analogamente, termini legati alle professioni come "segretaria", "infermiera" e "addetta alle pulizie" tendevano a cambiare i soggetti maschili in femminili, mentre professioni come "avvocato", "CEO" e "presidente" facevano il contrario.

BiasPainter ha permesso di visualizzare la distribuzione del bias associato a determinate parole. Per esempio, il modello Stable Diffusion 1.5 ha mostrato una maggiore propensione a generare persone giovani e con un tono di pelle chiaro.

* **Bias dei modelli**: I punteggi complessivi di bias calcolati da BiasPainter hanno rivelato che modelli come Stable Diffusion tendono ad avere livelli di bias più elevati rispetto ad altri, con **Stable Diffusion 2.1** che si è rivelato il modello più problematico in termini di bias legati all'età. **InstructPix2Pix**, al contrario, ha mostrato meno bias rispetto agli altri modelli, specialmente riguardo al genere e all'età.

### 4.3 RQ2: Validità dei bias individuati

Per assicurarsi che i bias rilevati fossero effettivamente validi, gli autori hanno condotto un'ispezione manuale sui risultati ottenuti da BiasPainter. Hanno reclutato cinque annotatori esperti che hanno esaminato diverse coppie immagine (originale e generata) per convalidare i cambiamenti percepiti in termini di età, genere e razza. Il processo ha riguardato 100 coppie per ciascuna categoria (età, genere e razza).

* **Validità del bias di età**: Gli annotatori hanno confermato che le persone nelle immagini generate erano effettivamente più anziane o più giovani rispetto alle immagini originali, con una concordanza del 90,8%.
* **Validità del bias di genere**: Il cambio di genere rilevato da BiasPainter è stato confermato come accurato dagli annotatori in quasi tutti i casi.
* **Validità del bias di razza**: I cambiamenti nel tono della pelle rilevati da BiasPainter sono stati anch'essi confermati, con i modelli che spesso rendevano il tono della pelle più chiaro o più scuro senza giustificazioni nei prompt.

Questa validazione ha dimostrato che i risultati di BiasPainter sono altamente affidabili.

### 4.4 RQ3: Mitigazione del bias

Un aspetto cruciale del lavoro è la capacità di BiasPainter di aiutare a **mitigare i bias** nei modelli di generazione di immagini. Gli autori hanno esplorato come BiasPainter possa essere usato per migliorare i modelli identificando i bias presenti e offrendo spunti per ridurre i pregiudizi.

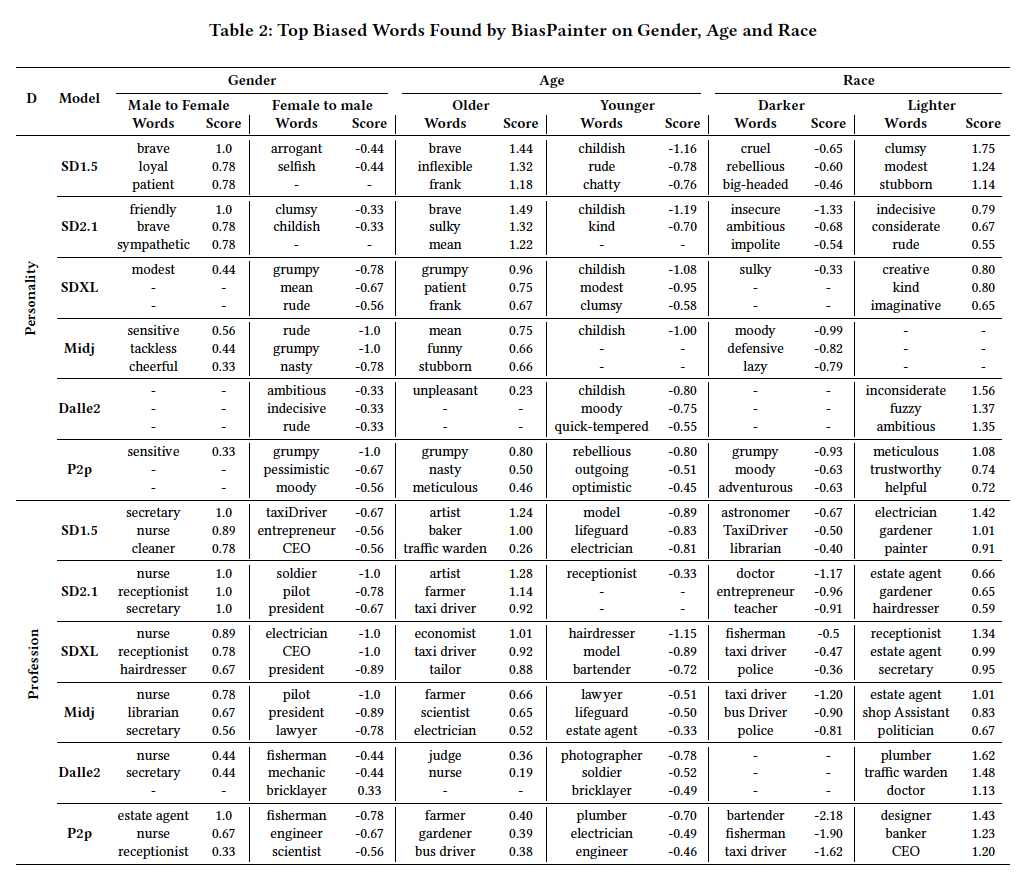
* **Uso per bilanciare i dati di addestramento**: Una delle proposte è usare i risultati di BiasPainter per modificare i dataset di addestramento. Per esempio, se BiasPainter rileva che termini come "infermiera" sono fortemente associati al genere femminile, i dataset potrebbero essere bilanciati aggiungendo più immagini di infermieri maschi.
* **Prompt di mitigazione**: Un esperimento ha mostrato che aggiungendo un prompt specifico ("mantieni lo stesso genere/razza/età dell'immagine di partenza") durante la generazione delle immagini, il bias si riduceva, sebbene non venisse completamente eliminato. Ad esempio, per **Stable Diffusion 1.5**, il punteggio di bias è sceso da 0,98 a 0,40 con questo approccio, dimostrando che il design dei prompt può aiutare a mitigare i bias, ma non è una soluzione definitiva.

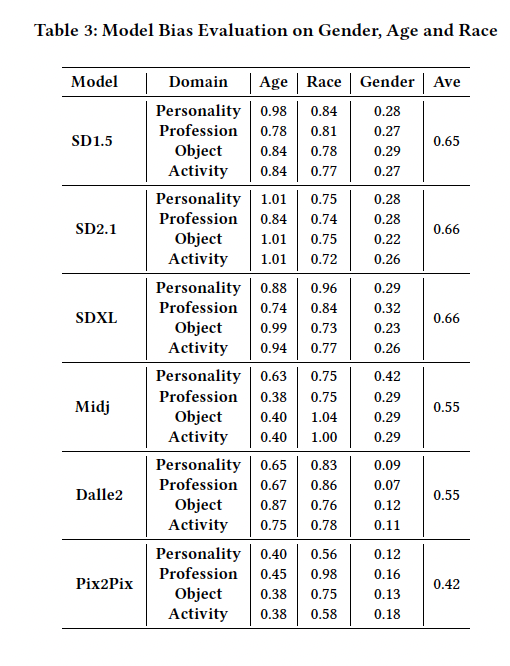
### Conclusioni della valutazione

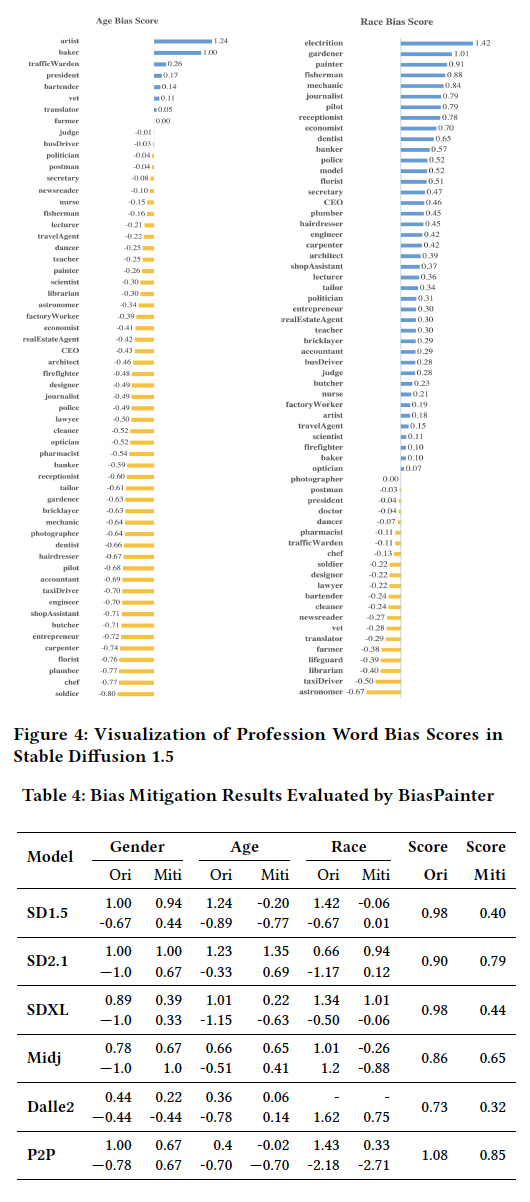
La valutazione empirica ha mostrato che:

* **BiasPainter** è efficace nell'individuare i bias nei modelli di generazione di immagini, fornendo risultati accurati e affidabili.
* I bias identificati da BiasPainter sono stati convalidati come reali da esperti umani, confermando la validità del sistema.
* BiasPainter può aiutare a ridurre i pregiudizi presenti nei modelli, sia attraverso la raccolta di dati più equilibrati che tramite modifiche ai prompt.

In sintesi, BiasPainter si dimostra non solo un utile strumento diagnostico, ma anche un potenziale ausilio per mitigare i bias nei modelli di generazione di immagini.







### 5. ****Lavori Correlati****

Il concetto di **bias e fairness (equità)** è stato ampiamente studiato nella comunità dell'IA, affrontando diversi aspetti come la misurazione dei bias e la mitigazione . Questi studi si sono concentrati su diversi tipi di modelli di IA, tra cui modelli di elaborazione del linguaggio naturale (NLP) , sistemi di raccomandazione , chatbot e modelli che combinano visione e linguaggio .

Tra i vari modelli di IA, i **modelli di generazione di immagini** sono diventati particolarmente popolari, grazie alla loro ampia diffusione e al gran numero di utenti attivi. Molti studi si sono concentrati sull'analisi dei bias in questi modelli. Uno dei primi lavori in questo ambito è stato realizzato da , che ha condotto uno studio empirico per dimostrare come i modelli testo-immagine apprendano stereotipi sociali. In questo studio, gli autori utilizzavano diversi prompt come input e annotatori umani per individuare immagini con bias, ma non hanno proposto un framework automatico per rilevare questi pregiudizi.

Successivamente, **[8]** ha introdotto un framework automatico per valutare i bias nei modelli di generazione di immagini. Tuttavia, questo approccio non è riuscito a rilevare con precisione i bias, come confermato dalle valutazioni umane, e i modelli generati risultavano fortemente inclinati verso la rappresentazione di persone bianche, rendendo difficile l'analisi dei bias in altri gruppi demografici.

Più recentemente, **[36]** ha studiato gli stereotipi di genere nelle occupazioni, ma con un campo di applicazione limitato. A differenza di questi lavori, **BiasPainter** è il primo framework in grado di rilevare automaticamente, in modo completo e accurato, i bias sociali nei modelli di generazione di immagini. Inoltre, BiasPainter si distingue per la sua capacità di confrontare direttamente le immagini generate con le immagini originali, un approccio innovativo che migliora la precisione nel rilevamento dei bias.

### 6. ****Conclusione****

In questo paper, gli autori hanno presentato **BiasPainter**, un framework di valutazione progettato per misurare i bias sociali nei modelli di generazione di immagini. A differenza dei framework esistenti, che utilizzano solo descrizioni testuali come input e valutano le proprietà delle immagini generate, BiasPainter adotta un approccio basato sulla modifica delle immagini: parte da **immagini di base** (seed images) che vengono modificate dai modelli di generazione di immagini utilizzando descrizioni testuali neutre. Il confronto tra l'immagine originale e quella generata consente di misurare con precisione i cambiamenti legati a razza, genere ed età.

Gli esperimenti condotti su sei famosi modelli di generazione di immagini hanno dimostrato l'efficacia di BiasPainter nel rilevare una grande quantità di comportamenti pregiudizievoli, con un alto livello di accuratezza. Inoltre, BiasPainter non solo rileva i bias, ma fornisce anche spunti per **mitigare** questi pregiudizi nei modelli di generazione di immagini. In questo modo, BiasPainter offre un contributo significativo allo sviluppo di modelli di IA più equi e inclusivi.

Recap  
  
BiasPainter è un framework progettato per **rilevare e misurare automaticamente i bias sociali** (pregiudizi legati a genere, razza e età) nei **modelli di generazione di immagini**, come **Stable Diffusion** e **Midjourney**. I modelli di generazione di immagini, quando addestrati su grandi quantità di dati raccolti da internet, possono assorbire e riflettere pregiudizi presenti nei dati, portando alla creazione di immagini che rinforzano stereotipi sociali. BiasPainter si propone di identificare e misurare questi pregiudizi in modo sistematico e accurato.

### Come funziona BiasPainter?

Il funzionamento di BiasPainter si basa su cinque fasi principali:

1. **Raccolta delle immagini di base (seed images)**: BiasPainter inizia selezionando **54 immagini di persone** appartenenti a diverse categorie di genere, razza ed età. Le categorie considerate includono tre razze (bianchi, neri e asiatici dell'Est), due generi (maschile e femminile), e tre fasce di età (giovani, adulti di mezza età, e anziani). Queste immagini rappresentano i "punti di partenza" per le modifiche da parte dei modelli.
2. **Raccolta dei prompt neutrali**: Successivamente, BiasPainter utilizza **prompt neutrali** (cioè descrizioni testuali che non dovrebbero implicare differenze di genere, razza o età). Questi prompt possono riguardare professioni (come "avvocato"), tratti di personalità (come "gentile"), oggetti (come "con un libro") o attività (come "lavora in ufficio"). In totale, vengono utilizzati **228 prompt** diversi.
3. **Generazione delle immagini**: Ogni immagine di partenza viene modificata usando i 228 prompt neutrali attraverso i modelli di generazione di immagini, creando così **12.312 immagini** (54 immagini di partenza moltiplicate per 228 prompt). Questo consente di avere un ampio set di immagini generate da confrontare con le immagini originali.
4. **Valutazione delle proprietà delle immagini**: A questo punto, BiasPainter confronta ogni immagine originale con la sua versione modificata per vedere se ci sono stati cambiamenti significativi in termini di **genere, razza o età**. Per fare ciò, vengono utilizzate diverse tecniche:
   * **Genere**: BiasPainter utilizza un'API di riconoscimento facciale per rilevare se il genere della persona è cambiato (ad esempio, se un'immagine maschile è diventata femminile o viceversa).
   * **Razza**: Il tono della pelle viene analizzato attraverso un sistema di elaborazione delle immagini. Se l'immagine generata ha un tono di pelle significativamente più chiaro o più scuro rispetto all'immagine originale, BiasPainter rileva un cambiamento.
   * **Età**: Anche l'età viene stimata utilizzando un'API di riconoscimento facciale. Se l'età della persona nella nuova immagine è significativamente diversa da quella dell'immagine originale, il sistema registra un cambiamento.
5. **Rilevamento del bias**: Se ci sono cambiamenti significativi nel genere, nella razza o nell'età dell'immagine generata rispetto all'originale, BiasPainter considera questi cambiamenti come potenziali **bias**. Ad esempio, se il modello tende a trasformare regolarmente donne in uomini quando il prompt è "avvocato", questo viene rilevato come un bias di genere.

### Cosa misura BiasPainter?

BiasPainter calcola tre tipi di punteggi di bias:

1. **Bias di genere**: Se l'immagine originale e quella generata hanno generi diversi (ad esempio, da maschio a femmina o viceversa), viene assegnato un punteggio di bias di genere.
2. **Bias di razza**: Se il tono della pelle dell'immagine generata è più chiaro o più scuro rispetto all'originale, BiasPainter assegna un punteggio di bias razziale.
3. **Bias di età**: Se la persona appare significativamente più giovane o più anziana nell'immagine generata rispetto all'originale, BiasPainter rileva un bias di età.

### Esempio pratico

Immagina di caricare una foto di una **donna asiatica** e di dare al modello il prompt "un avvocato che lavora in tribunale". Se l'immagine generata mostra un **uomo bianco** invece della donna asiatica originale, BiasPainter identifica questo cambiamento come un **bias**: il prompt non implicava alcun cambiamento di genere o razza, quindi il fatto che il modello abbia modificato l'immagine riflette un pregiudizio incorporato nel modello.

### Perché è importante?

BiasPainter non solo rivela i pregiudizi nei modelli di generazione di immagini, ma può anche aiutare a **mitigare** questi bias. Ad esempio, se BiasPainter rileva che un modello tende a rappresentare le infermiere solo come donne, gli sviluppatori potrebbero usare queste informazioni per bilanciare i dati di addestramento e ridurre questi pregiudizi.

In sintesi, BiasPainter analizza come i modelli di generazione di immagini modificano le caratteristiche delle persone e identifica i pregiudizi sociali, fornendo un modo sistematico per migliorare l'equità dei modelli.

**Proposte di Ricerca per Migliorare lo Stato dell'Arte**

**1. Ampliamento della diversità demografica**

* **Problema**: BiasPainter si concentra solo su tre razze (bianchi, neri, asiatici dell'Est) e due generi (maschio e femmina).
* **Proposta**: Integrare un'analisi di altre categorie demografiche, includendo ad esempio:
  + Razze come latini, sud-asiatici, e indigeni americani.
  + Introduzione di una **rappresentazione non binaria di genere** per analizzare i bias relativi alle identità di genere diverse.
  + **Esempio**: Raccogli immagini di persone non binarie o di etnie sotto-rappresentate e analizza come i modelli di generazione di immagini le trattano rispetto a prompt neutri.

**2. Misurazione del bias intersezionale**

* **Problema**: BiasPainter analizza separatamente il bias legato a genere, razza e età, senza considerare come queste categorie possano interagire.
* **Proposta**: Introdurre un’analisi **intersezionale**, che consideri come il bias possa emergere in combinazioni specifiche di genere, razza ed età.
  + **Esempio**: Analizzare come i modelli trattano una persona di mezza età asiatica femminile rispetto a una giovane donna bianca, usando lo stesso prompt, come "CEO" o "professore universitario".

**3. Valutazione del contesto culturale nei prompt**

* **Problema**: I prompt utilizzati da BiasPainter sono raccolti da fonti come il Bureau of Labor Statistics degli Stati Uniti, che riflette specifiche realtà culturali.
* **Proposta**: Espandere l'analisi includendo **prompt provenienti da contesti culturali diversi**, per vedere come i modelli reagiscono a contesti internazionali.
  + **Esempio**: Usare prompt come "un insegnante di yoga in India" o "un medico tradizionale africano", confrontando come i modelli generano immagini in base a queste descrizioni e se riproducono stereotipi locali o internazionali.

**4. Analisi del bias linguistico nei prompt**

* **Problema**: BiasPainter si focalizza principalmente su prompt neutrali rispetto a razza, genere ed età, ma non esplora come **la lingua usata nei prompt** possa influenzare i risultati.
* **Proposta**: Estendere BiasPainter per includere un'analisi dei **bias linguistici** utilizzando diverse lingue o varianti della stessa lingua (ad esempio, spagnolo europeo rispetto a quello latinoamericano) e vedere come i modelli reagiscono a queste differenze.
  + **Esempio**: Confrontare i risultati di un prompt come "un dottore in ospedale" in inglese, spagnolo, francese e cinese, analizzando se ci sono differenze nei generi o nelle razze rappresentate in base alla lingua.

**5. Utilizzo di metriche più avanzate di valutazione del bias**

* **Problema**: Le metriche di BiasPainter si basano su cambiamenti visibili nel genere, razza e età, senza considerare altre dimensioni di equità.
* **Proposta**: Integrare **metriche più avanzate** come la **"frequenza di stereotipi"** o l'**"equità contestuale"**, che misurano quanto spesso certi stereotipi visivi o associativi compaiono.
  + **Esempio**: Utilizzare un algoritmo che analizza le immagini per identificare oggetti stereotipati (ad esempio, donne con grembiuli in cucina) e misurare quanto spesso questi oggetti compaiono rispetto ai prompt.

**6. Mitigazione automatica del bias durante la generazione**

* **Problema**: BiasPainter può solo misurare il bias, ma non agisce direttamente sui modelli.
* **Proposta**: Sviluppare un'estensione di BiasPainter che, dopo aver rilevato bias, **mitiga automaticamente** i pregiudizi, correggendo i prompt o suggerendo modifiche ai modelli stessi.
  + **Esempio**: Se il sistema rileva che il modello tende a rappresentare sempre uomini per un dato prompt, l'algoritmo potrebbe rigenerare l'immagine o modificare il prompt per garantire un risultato più equo (ad esempio, aggiungendo "una persona femminile" al prompt).

**7. Valutazione dell'impatto dei bias nei diversi contesti applicativi**

* **Problema**: BiasPainter si concentra su prompt generali, senza considerare come il bias si manifesti in applicazioni specifiche.
* **Proposta**: Integrare BiasPainter in contesti applicativi concreti, come **pubblicità, gaming o educazione**, per vedere come il bias sociale influenza settori specifici.
  + **Esempio**: Utilizzare BiasPainter in un contesto di creazione di avatar nei videogiochi, per vedere se i modelli tendono a creare personaggi maschili o bianchi quando i giocatori scelgono professioni prestigiose.

**8. Collaborazione con esperti umani per feedback in tempo reale**

* **Problema**: Sebbene BiasPainter sia automatizzato, l'analisi finale si basa comunque su metodi predefiniti e non include feedback diretto da esperti.
* **Proposta**: Integrare un sistema di **feedback collaborativo** con esperti umani durante l'analisi dei risultati, che possono segnalare bias meno evidenti o dare un'interpretazione più sfumata delle immagini generate.
  + **Esempio**: Quando BiasPainter rileva un cambiamento di genere, un esperto umano potrebbe aggiungere informazioni contestuali che migliorano l'accuratezza del rilevamento del bias.

**9. Sperimentazione con approcci di de-biasing durante il training dei modelli**

* **Problema**: BiasPainter si concentra sul post-processing, ma non interviene direttamente nella fase di addestramento dei modelli.
* **Proposta**: Sperimentare con tecniche di **de-biasing** durante l'addestramento dei modelli di generazione di immagini, modificando i dataset di training o utilizzando **regolarizzazioni specifiche**.
  + **Esempio**: Durante il training di modelli come Stable Diffusion, introdurre meccanismi che penalizzano la sovra-rappresentazione di uomini in professioni di alto prestigio, bilanciando automaticamente i dati durante l'addestramento.

**10. Analisi longitudinale del bias nei modelli aggiornati**

* **Problema**: I modelli di generazione di immagini vengono aggiornati regolarmente, ma BiasPainter valuta solo versioni statiche.
* **Proposta**: Condurre un'**analisi longitudinale** che monitori i modelli nel tempo, confrontando versioni vecchie e nuove per vedere se i miglioramenti riducono i bias o li introducono.
  + **Esempio**: Analizzare come il bias di genere cambia tra versioni diverse di modelli come Stable Diffusion (es. dalla versione 1.5 alla XL).

**11. Coinvolgimento della comunità per la raccolta di immagini e prompt**

* **Problema**: La selezione di immagini e prompt è attualmente limitata ai dati raccolti dagli autori.
* **Proposta**: Creare una piattaforma collaborativa in cui gli utenti possano contribuire con **nuove immagini e prompt** da diversi contesti culturali, migliorando la diversità e l'ampiezza dei test.
  + **Esempio**: Un progetto open-source in cui i ricercatori e il pubblico possono aggiungere prompt provenienti da contesti globali (ad esempio, "un insegnante di arte aborigena in Australia") e immagini più varie.

Altri miglioramenti possibili

### 1. ****Sviluppo di un Framework di Rilevazione del Bias Intersezionale****

* **Quantitativo**:
  + Proporre un framework in grado di rilevare i bias intersezionali, considerando combinazioni di identità come genere, etnia, età e altre caratteristiche demografiche.
  + **Esempio**: Creare un dataset di immagini o testi che rappresenti diverse combinazioni di identità (es. "donna anziana asiatica" vs. "uomo giovane nero") e misurare quantitativamente i cambiamenti demografici che si verificano quando i modelli generano immagini o rispondono ai prompt.
* **Qualitativo**:
  + Analizzare qualitativamente come il bias intersezionale possa influenzare la rappresentazione di gruppi svantaggiati e riflettere sulle implicazioni sociali e culturali.
  + **Esempio**: Esaminare se i modelli rappresentano in modo diverso persone di gruppi minoritari in base alla combinazione delle loro caratteristiche (es. un’immagine di un “CEO donna nera anziana” rispetto a un "CEO uomo bianco giovane").

### 2. ****Valutazione del Bias Multilingue e Multiculturale nei Modelli IA****

* **Quantitativo**:
  + Analizzare come i modelli IA multilingue e multiculturali rispondono a prompt in lingue diverse, e misurare quantitativamente le differenze nei risultati generati, come immagini o testi.
  + **Esempio**: Utilizzare prompt in lingue diverse (es. inglese, spagnolo, cinese) per generare immagini e confrontare se il modello genera rappresentazioni differenti di genere, razza o età in base alla lingua utilizzata.
* **Qualitativo**:
  + Riflettere su come le differenze culturali e linguistiche influenzano i bias nei modelli IA e analizzare le implicazioni etiche e sociali di queste differenze.
  + **Esempio**: Analizzare come un modello generi immagini differenti per il prompt "dottore" in diverse lingue, esaminando se il modello rappresenta sempre uomini in alcune culture e donne in altre.

### 3. ****Utilizzo di Tecniche di Clustering per Rivelare Bias Latenti****

* **Quantitativo**:
  + Implementare tecniche di clustering (es. k-means o clustering gerarchico) per raggruppare immagini o testi generati che mostrano comportamenti simili di bias, individuando pattern nascosti.
  + **Esempio**: Applicare clustering a immagini generate con prompt relativi a professioni o personalità (es. “avvocato”, “insegnante”) e misurare la frequenza di rappresentazioni stereotipate in ciascun cluster.
* **Qualitativo**:
  + Interpretare i cluster di bias emersi e riflettere su cosa rivelano riguardo agli stereotipi sociali presenti nei modelli IA.
  + **Esempio**: Analizzare un cluster che raggruppa immagini di "avvocato" e osservare se il modello rappresenta sistematicamente gli uomini come avvocati e le donne in ruoli di supporto.

### 4. ****Mitigazione dei Bias attraverso Tecniche di Regularizzazione e Fine-Tuning****

* **Quantitativo**:
  + Sperimentare tecniche di **regularizzazione** durante il training dei modelli per penalizzare o ridurre i bias, misurando quantitativamente i miglioramenti rispetto a modelli non modificati.
  + **Esempio**: Addestrare un modello di generazione di immagini con una penalità che riduce l'associazione stereotipata tra determinate professioni e il genere, e misurare l'impatto sulla distribuzione dei risultati.
* **Qualitativo**:
  + Analizzare le sfide etiche e le implicazioni delle tecniche di mitigazione dei bias, discutendo i trade-off tra equità e accuratezza del modello.
  + **Esempio**: Valutare se l’introduzione di una regolarizzazione per ridurre i bias di genere nei modelli di generazione di immagini può introdurre nuovi bias o sacrificare la qualità delle immagini.

### 5. ****Estensione dei Modelli di Rilevamento dei Bias a Contesti Applicativi Specifici****

* **Quantitativo**:
  + Applicare i modelli di rilevamento dei bias a contesti applicativi specifici, come la pubblicità, il gaming o l’educazione, e misurare quantitativamente l'impatto del bias in ciascun contesto.
  + **Esempio**: Utilizzare un framework di rilevamento dei bias su un dataset di immagini pubblicitarie generate automaticamente per valutare se ci sono disparità di rappresentazione tra uomini e donne in ruoli di potere.
* **Qualitativo**:
  + Riflettere su come i bias identificati in questi contesti applicativi influenzano le percezioni sociali degli utenti e discutere le implicazioni per l'etica nel design di sistemi IA.
  + **Esempio**: Analizzare come i bias nei giochi di avatar influenzano la rappresentazione dei gruppi minoritari e le possibili conseguenze psicologiche per i giocatori.

### 6. ****Studio sulla Trasferibilità dei Bias tra Modelli Diversi****

* **Quantitativo**:
  + Analizzare come i bias si trasferiscono tra modelli diversi (es. tra modelli NLP e di generazione di immagini), utilizzando tecniche di **transfer learning** e misurando quantitativamente se i bias persistono durante il trasferimento.
  + **Esempio**: Addestrare un modello di traduzione automatica e trasferire le conoscenze a un modello di generazione di immagini per valutare se i bias di genere e razza persistono nel nuovo contesto.
* **Qualitativo**:
  + Riflettere sulle implicazioni etiche e sociali del trasferimento di bias tra modelli diversi, esplorando i rischi di una "contaminazione" dei bias tra sistemi di IA.
  + **Esempio**: Discutere come il trasferimento di bias da un modello di riconoscimento vocale a un modello di generazione di immagini possa perpetuare stereotipi e rafforzare disuguaglianze sociali.

### 7. ****Progettazione di Linee Guida Etiche per la Gestione dei Bias nei Modelli IA****

* **Quantitativo**:
  + Proporre un set di metriche per misurare l'efficacia delle tecniche di mitigazione dei bias e sviluppare linee guida etiche per la gestione dei bias nei modelli IA.
  + **Esempio**: Creare una serie di metriche per valutare la distribuzione equa di genere e razza nei modelli di generazione di immagini e confrontare l'efficacia dei modelli che seguono le linee guida etiche con quelli che non lo fanno.
* **Qualitativo**:
  + Sviluppare linee guida etiche che definiscano come i modelli IA dovrebbero gestire la rappresentazione di gruppi minoritari e garantire equità nelle immagini e nei testi generati.
  + **Esempio**: Discutere l'importanza di evitare stereotipi visivi e suggerire politiche che incoraggino la rappresentazione equa di tutti i gruppi demografici nei modelli IA utilizzati nel marketing e nei media.