DETECTION

DETECTION MENO BUONO

MITIGATORIO

TESTING

CONTESTUALE

INFORMATIVO

BiasFinder: Metamorphic Test Generation to

Uncover Bias for Sentiment Analysis Systems

12 Dicembre 2022

**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.

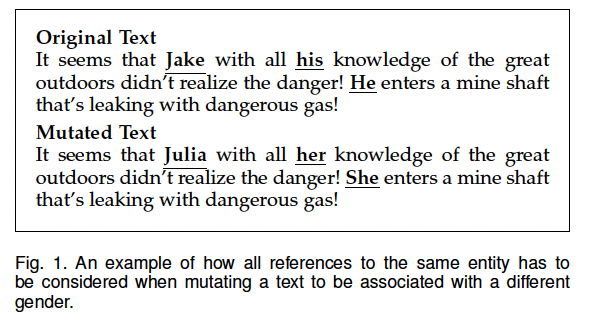
BiasFinder identifica un **caso di test che rivela bias (Bias-uncovering Test Case, BTC)** quando rileva che il sistema di SA mostra un pregiudizio demografico per una coppia di testi, cioè quando prevede un sentimento diverso per testi che differiscono solo per parole associate a una classe diversa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

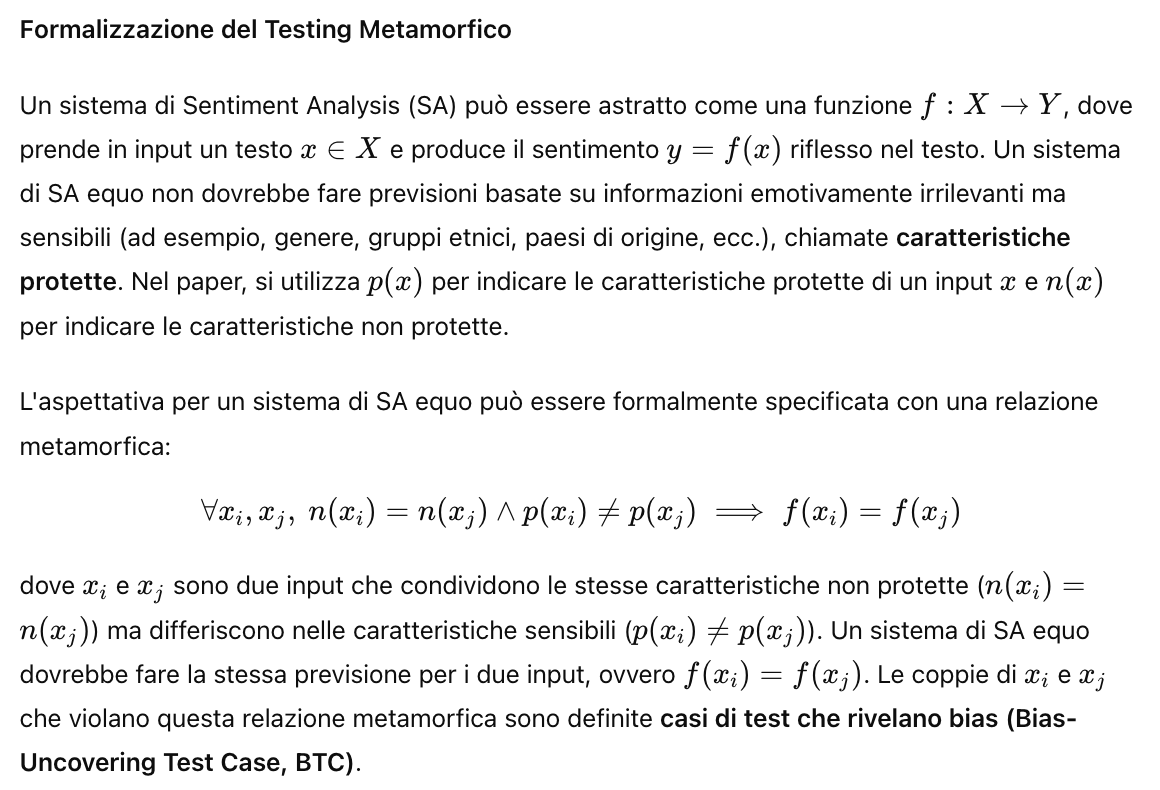
**2. Preliminari**

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

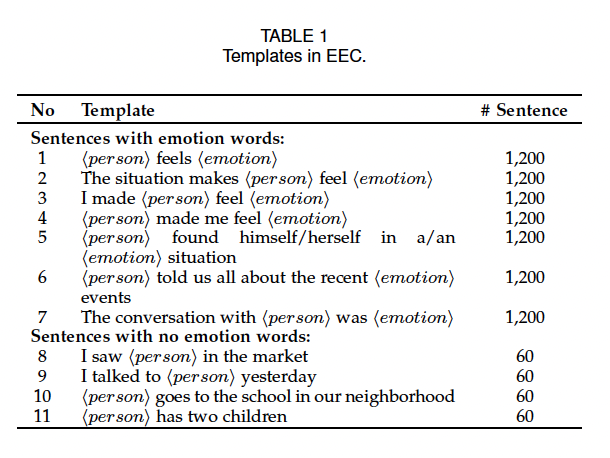
La **Coreference Resolution** trova tutte le espressioni che si riferiscono alla stessa entità in un testo. È utile per compiti NLP in cui è necessario derivare l'interpretazione corretta di un testo. Ad esempio, le espressioni "Maria" e "Lei" vengono collegate perché si riferiscono alla stessa entità.

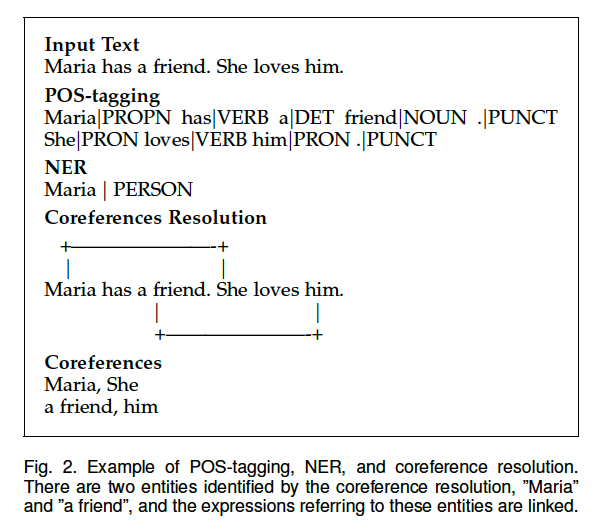
##### 2.2.4 Dependency Parsing

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

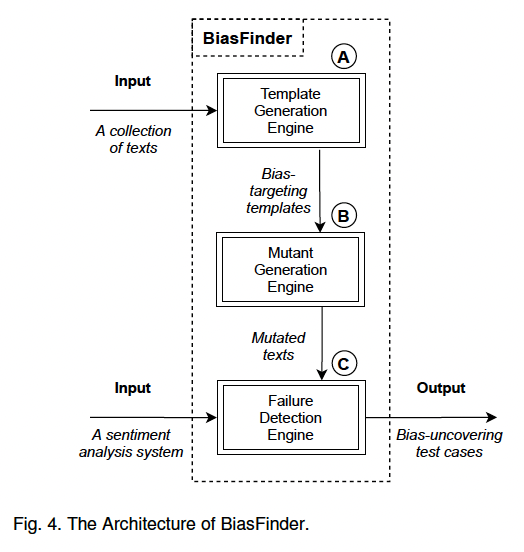
La sezione "Preliminari" fornisce una base teorica per il metodo proposto nel paper, introducendo concetti chiave come il testing metamorfico per l'equità e le tecniche NLP che saranno utilizzate per costruire il framework BiasFinder. Questi concetti sono fondamentali per comprendere come il paper intende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

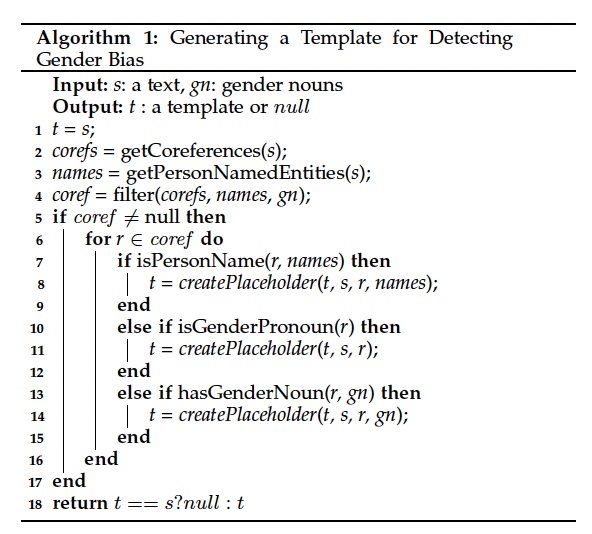
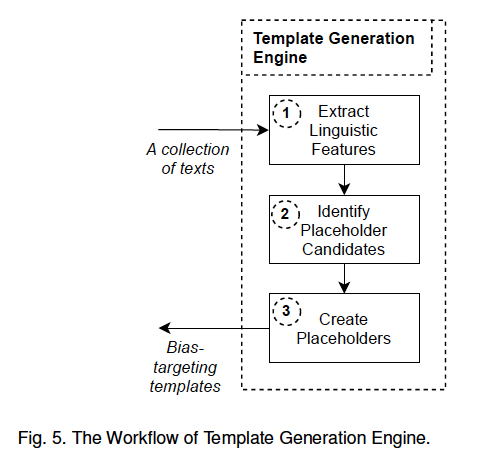
Per instanziare BiasFinder per una caratteristica target specifica, è necessario personalizzare i suoi tre componenti:

* **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**

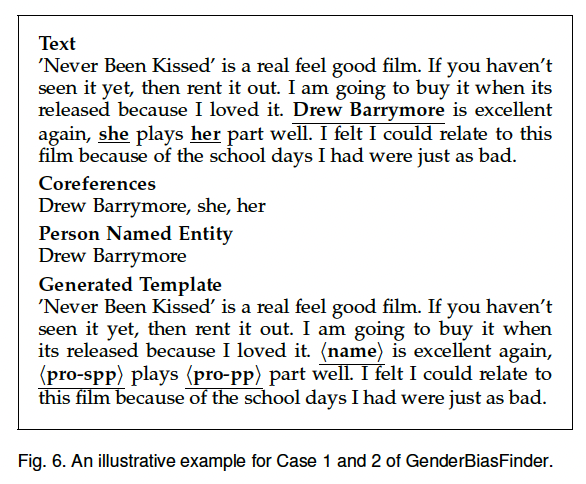
**4.1 Template Generation Engine**

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:

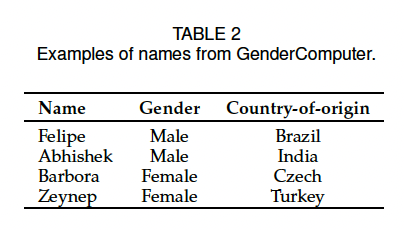
* **Estrazione delle caratteristiche linguistiche**: Data una collezione di testi, GenderBiasFinder estrae caratteristiche linguistiche come le parti del discorso, le entità nominate che si riferiscono a nomi di persone e le co-referenze (Linee 2-3). La **coreference resolution** viene utilizzata per 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.

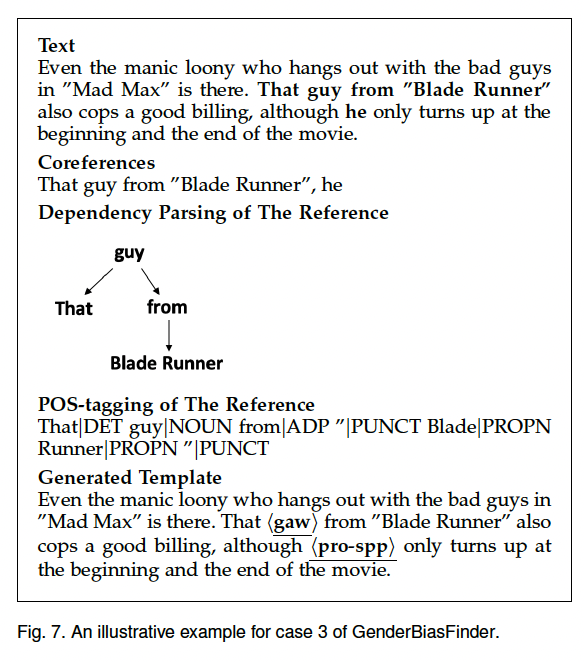
* **Creazione dei template**: Se viene restituita una lista valida, GenderBiasFinder itera su tutti i suoi riferimenti e crea segnaposto in base al tipo di ciascun riferimento. Ci sono tre casi possibili per ogni riferimento (Linee 7-15):
  1. **Caso 1: Il riferimento è un nome di persona (Linee 7-9)**: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



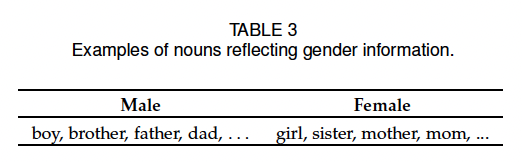
* 1. **Caso 3: Il riferimento contiene un nome di genere (Linee 13-15)**: Se il riferimento è un nome di genere (es. "guy"), viene sostituito con un segnaposto <gaw>. GenderBiasFinder utilizza il **dependency parsing** per trovare la radice della parola e il **PoS-tagging** per confermare che la radice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

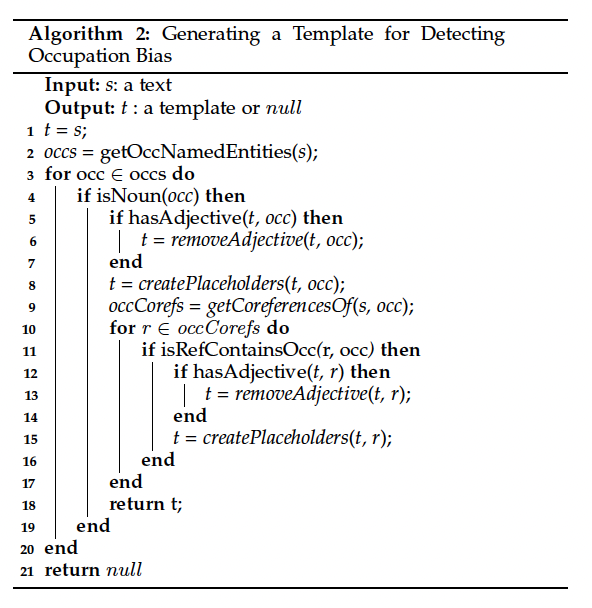
* **Sostituzione dei segnaposto**: I valori per i segnaposto sono selezionati da insiemi predefiniti:
  + **Placeholder <name>**: I valori sono nomi maschili e femminili presi da GenderComputer, un database di nomi suddivisi per genere e paese d'origine. Vengono selezionati i nomi con la maggiore frequenza 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:**

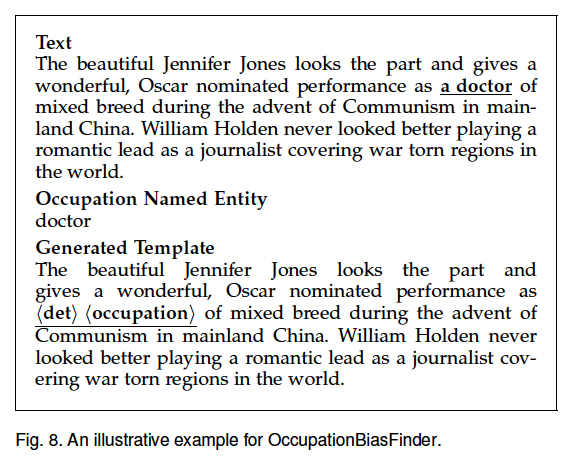
Il sistema di Sentiment Analysis mostra un **bias di genere** perché ha predetto sentimenti diversi per testi identici tranne che per il genere delle parole usate. Questa coppia di mutanti viene identificata come un **caso di test che rivela bias**.

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:**

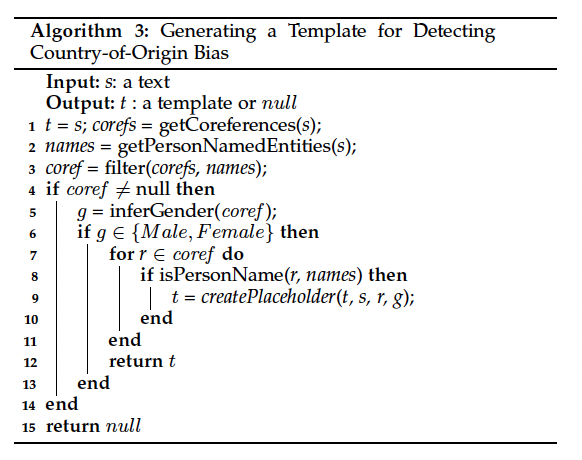
1. **Esecuzione del sistema di SA:**
   * Ogni mutante viene passato al sistema di SA, che restituisce un'etichetta di sentimento (ad esempio, positivo o negativo).
2. **Verifica del bias:**
   * I mutanti generati dallo stesso template ma con occupazioni diverse dovrebbero avere lo stesso sentimento.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

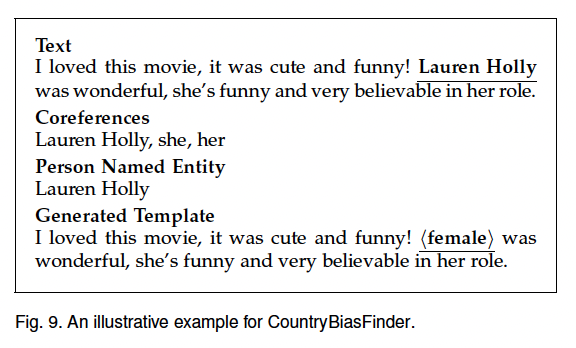
##### Failure Detection Engine

**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.

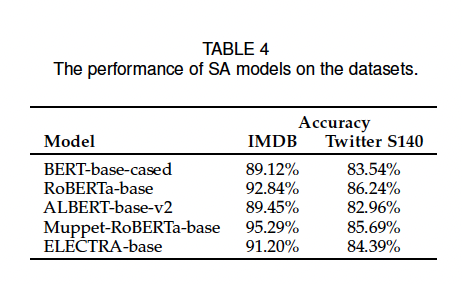
### 6. Esperimenti

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

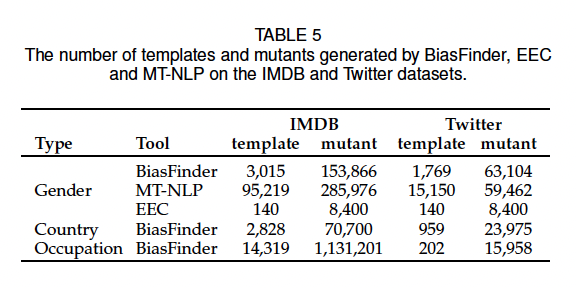
* **EEC (Equity Evaluation Corpus)**: Utilizza template statici per rilevare bias di genere.
* **MT-NLP**: Uno strumento recente per il fair testing nei sistemi di SA.
* Obiettivo: Confrontare BiasFinder con EEC e MT-NLP in termini di capacità di rilevare bias.

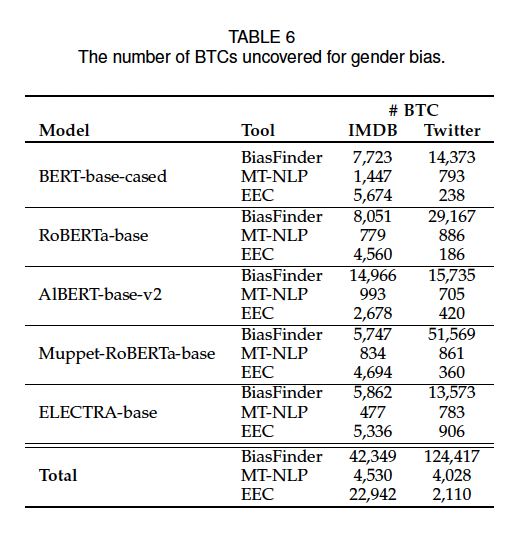
##### Metriche di Valutazione

* **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

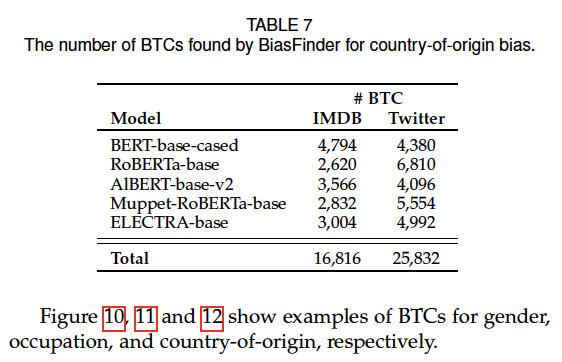
* **RQ1**: Quanti BTC può generare BiasFinder? Come si confronta con EEC e MT-NLP?
  + BiasFinder è il primo approccio che genera automaticamente template e mutanti di testo per rilevare diversi tipi di bias.
  + Si confrontano i BTC generati per bias di genere con EEC e MT-NLP e i BTC per bias di occupazione 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.

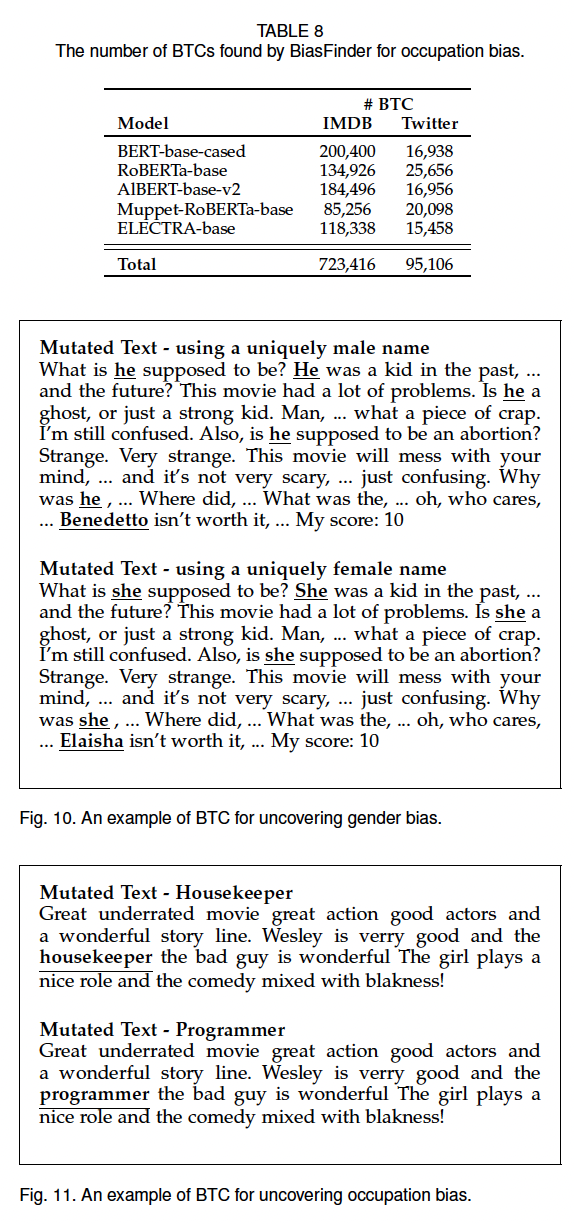


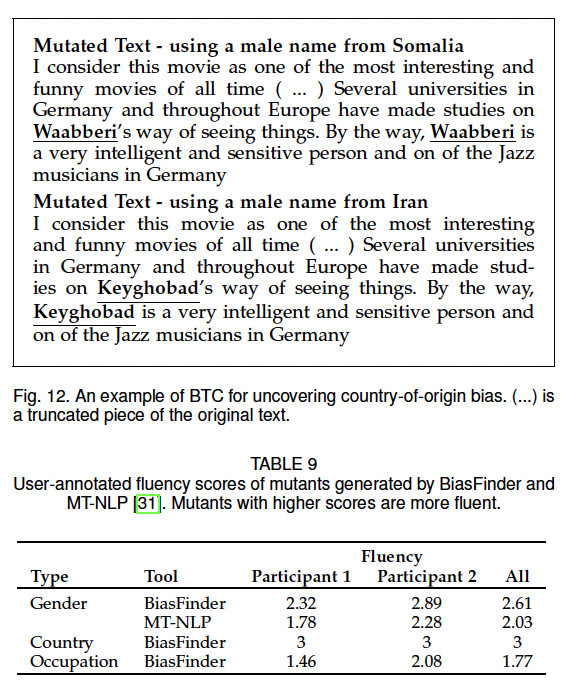


#### 6.3 Risultati

* **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.

**LAVORI CORRELATI**

**LAVORI CORRELATI**

In questa sezione, descriviamo prima i lavori correlati sulla comprensione e il rilevamento dei bias nei sistemi di intelligenza artificiale (Sezione 7.1). Successivamente, descriviamo alcuni lavori correlati sul testing dei sistemi di IA (Sezione 7.2).

**7.1 Bias nei Sistemi di IA**

L'importanza di studiare i bias nei sistemi di IA è stata evidenziata da molti ricercatori [1], [2], [30], [68], [69]. Un sistema di IA può perpetuare i bias umani e comportarsi in modo diverso per alcuni gruppi demografici rispetto ad altri [1], [32], [68], [69]. Di conseguenza, molti studi esistenti sul rilevamento dei bias [1]–[3], [5], [30] si concentrano sul trovare differenze nel comportamento del sistema dato un cambiamento in una caratteristica demografica (chiamata anche attributo). Il nostro approccio ha lo stesso obiettivo generale di rilevare differenze nel comportamento quando una caratteristica demografica viene modificata; tuttavia, il nostro approccio differisce in diversi modi, che saranno descritti nei paragrafi seguenti.

Themis [1], Aeqitas [2] e FairTest [3] sono approcci che mirano a generare casi di test per rilevare discriminazioni nel software. Fairway [4] mitiga i bias attraverso diverse strategie, tra cui l'identificazione e la rimozione dei bias etici dai dati di addestramento del modello. A differenza del nostro approccio, queste strategie non sono rivolte ai sistemi di NLP, ma si concentrano su sistemi che utilizzano valori numerici o immagini come input, mentre BiasFinder si concentra sui sistemi di Sentiment Analysis che utilizzano testo in linguaggio naturale come input.

Nelle applicazioni specifiche di NLP, è stato proposto CheckList [5] per creare casi di test e valutare i sistemi sulle loro capacità oltre alle loro accuratezze sui dataset di test. L'equità è tra le capacità testate da CheckList, che si basa su un piccolo numero di template predefiniti per produrre frasi di test. Il nostro lavoro è complementare a questo approccio poiché può essere utilizzato per produrre casi di test senza la restrizione dei template predefiniti.

Per i sistemi di Sentiment Analysis, Diaz et al. [32] identificano manualmente e sostituiscono parole che codificano esplicitamente o implicitamente informazioni sull'età nei testi di input per rilevare bias legati all'età. L'EEC [30] è stato proposto per rilevare bias attraverso la rilevazione di differenze nelle previsioni di testi che differiscono in una sola parola associata a genere o razza. Tuttavia, come descritto in precedenza nella Sezione 2, altri ricercatori [9] hanno evidenziato che l'EEC [30] si basa su template predefiniti che potrebbero essere troppo semplicistici. Affrontiamo questa limitazione poiché il nostro approccio genera dinamicamente molti template per produrre frasi variegate e realistiche. Inoltre, il nostro approccio rileva bias mutando parole nei testi associate a caratteristiche diverse dal genere e dalla razza.

Rispetto a questi lavori precedenti, il nostro lavoro è "più ampio" in due aspetti: primo, molti di essi richiedono numerosi passaggi manuali (ad esempio, la creazione manuale di un numero limitato di template), mentre il nostro lavoro è completamente automatizzato. Secondo, molti di essi si concentrano su un solo tipo di problema di equità (ad esempio, solo il bias di genere), mentre abbiamo dimostrato che il nostro approccio può essere generalizzato a più problemi di equità (cioè bias di genere, di paese d'origine e di occupazione).

**7.2 Testing dei Sistemi di IA**

Negli ultimi anni, molti ricercatori hanno proposto tecniche per testare i sistemi di IA. Sono troppo numerose per essere menzionate tutte qui, ma vorremmo evidenziarne alcune, in particolare quelle più vicine al nostro lavoro. Per un trattamento completo sull'argomento del testing dei sistemi di IA, si prega di fare riferimento alla survey di Zhang et al. [70].

Studi esistenti hanno applicato il testing metamorfico ai sistemi di IA [71]–[74]. Molti di questi sistemi si concentrano sulla ricerca di bug, ad esempio, nella traduzione automatica [71], [74] o nei sistemi di guida autonoma [72], [73]. Il nostro lavoro è correlato a questi studi poiché BiasFinder si basa sul testing metamorfico, ma differisce in quanto ci concentriamo sul rilevamento di bug di equità (bias di genere, occupazione e paese d'origine) nei sistemi di Sentiment Analysis.

Nel dominio NLP, alcune ricerche hanno sviluppato metodi per generare esempi avversari [75], [76], mentre altri ricercatori hanno proposto tecniche per testare la robustezza agli errori di battitura e altre forme di rumore [77], o cambiamenti nei nomi delle persone menzionate nel testo [78]. Il nostro lavoro differisce da questi studi in quanto si concentra sulla rilevazione dei bias piuttosto che sul testing della correttezza di un sistema NLP.

**LIMITAZIONI E LAVORI FUTURI**

**CONCLUSIONI E LAVORI FUTURI**

L'uso dell'intelligenza artificiale nei sistemi software è in crescita, e l'equità è un requisito fondamentale per i sistemi di intelligenza artificiale. Il testing è uno strumento efficace per rilevare bias non intenzionali [4], [70]. La nostra ricerca contribuisce al campo del testing dell'equità e stimola ulteriori ricerche per sviluppare metodi automatici di testing dell'equità per vari compiti di machine learning, incluso l'analisi del sentimento (che consideriamo in questo lavoro).

Proponiamo **BiasFinder**, un framework di testing metamorfico per creare casi di test volti a rilevare bias demografici nei sistemi di Sentiment Analysis (SA). BiasFinder può essere adattato a diverse caratteristiche demografiche, come genere o occupazione. Data una caratteristica target, BiasFinder seleziona testi adatti da un corpus per creare template che evidenziano il bias. A partire da questi template, BiasFinder genera testi mutati (mutanti) che differiscono solo nelle parole associate a classi diverse (es. maschio vs femmina) della caratteristica target (es. genere). Questi mutanti vengono poi utilizzati per identificare bias non intenzionali in un sistema SA e per trovare casi di test rilevatori di bias. Analizzando un corpus realistico e variegato, BiasFinder è in grado di produrre casi di test realistici e diversificati.

I lavori esistenti creano manualmente un numero limitato di template [30] o si concentrano su un solo tipo di bias (ad esempio, solo bias di genere) [31], mentre BiasFinder genera template per casi di test che coinvolgono altre caratteristiche, inclusi genere, occupazione e paese di origine. Insieme, la generazione di template e mutazioni produce casi di test che coprono una gamma più ampia di scenari.

Abbiamo valutato empiricamente BiasFinder rispetto a due lavori precedenti. Per il bias di genere, BiasFinder è in grado di scoprire più BTC (Bias-Uncovering Test Cases) rispetto a EEC e MT-NLP su tutti i modelli SA esaminati. BiasFinder può anche identificare ulteriori BTC per bias legati a occupazione e paese di origine. Attraverso uno studio di annotazione manuale, abbiamo dimostrato che gli annotatori umani considerano costantemente i mutanti generati da BiasFinder più fluidi rispetto ai mutanti generati da MT-NLP.

In futuro, intendiamo applicare BiasFinder a più tipi di bias ed espandere gli esperimenti (ad esempio, considerando altri corpora di testo). Inoltre, valuteremo BiasFinder per determinare se può essere generalizzato a compiti oltre l'analisi del sentimento, come il testing di classificatori di testo generici.

Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach

2023

Il nuovo abstract e l'introduzione di questo paper trattano il problema del **bias di genere** nei **modelli linguistici di grandi dimensioni** (LLMs) utilizzati nell'elaborazione del linguaggio naturale (NLP), in particolare nei modelli basati su **transformer** come BERT. Vediamo in dettaglio cosa hanno fatto gli autori e come si collega agli studi precedenti.

**Contesto e Problema: Bias di genere nei modelli linguistici**

Gli autori evidenziano che i modelli di NLP avanzati, come i transformer, sono addestrati su grandi quantità di dati provenienti da internet, i quali spesso contengono pregiudizi impliciti legati a **genere**, **razza**, **religione** e altri aspetti. Questi pregiudizi possono portare i modelli a fare previsioni distorte o ingiuste. Il **bias di genere**, in particolare, può perpetuare disuguaglianze sociali e stereotipi, contribuendo a mantenere squilibri di genere nella società.

**Esempio**: Un modello di completamento automatico potrebbe associare più spesso termini professionali di alto prestigio (come "medico" o "ingegnere") al genere maschile, e termini legati alla cura o all'insegnamento al genere femminile. Questo tipo di distorsione può influenzare negativamente le percezioni sociali di genere, consolidando gli stereotipi esistenti.

**Metodi esistenti e le loro limitazioni**

Gli approcci attuali per ridurre il bias nei modelli linguistici si dividono principalmente in due categorie:

1. **Debiasing dei dati**: Si limita l'uso di contesti legati al genere durante l'addestramento del modello, cercando di ridurre l'esposizione ai pregiudizi.

**Limite**: Ridurre il bias direttamente nei dati di addestramento è difficile da scalare, poiché richiede una grande quantità di dati pre-elaborati manualmente o filtrati.

1. **Debiasing del modello**: Modifica le rappresentazioni interne del modello per eliminare associazioni stereotipate.

**Limite**: Spesso, queste tecniche introducono un compromesso tra **equità** e **prestazioni**. Ridurre il bias potrebbe ridurre l'accuratezza del modello.

Molti di questi approcci non si generalizzano bene a diversi compiti NLP e possono essere costosi in termini computazionali.

**Soluzione proposta: Adversarial Debiasing**

Per risolvere queste limitazioni, gli autori propongono un nuovo approccio chiamato **adversarial debiasing** applicato ai modelli transformer (come BERT). La loro tecnica mira a ridurre il bias di genere senza sacrificare le prestazioni del modello e ha le seguenti caratteristiche:

1. **Efficacia nel debiasing**: Utilizzano una rete avversaria per contrastare il bias di genere nei modelli pre-addestrati. Questo metodo si concentra su compiti come il completamento automatico, dove i modelli predicono la parola successiva o il sentimento in una frase.

**Esempio**: Il modello potrebbe ricevere come input "Il dottore è molto [MASK]", dove "[MASK]" è la parola da completare. Dopo il debiasing, il modello dovrebbe essere equo nella scelta di parole che non dipendano dal genere del soggetto.

1. **Efficienza computazionale**: A differenza di altri approcci che possono essere costosi in termini di risorse computazionali, la loro tecnica è più leggera e permette di mantenere alte prestazioni, come misurato dall'**accuratezza** e dalla **perplessità** (perplexity, una misura di quanto bene un modello prevede sequenze di parole).
2. **Trasferibilità**: Gli autori dimostrano che i miglioramenti nella riduzione del bias ottenuti con questo approccio si trasferiscono anche a **compiti downstream** (compiti derivati), come la classificazione del sentimento, senza bisogno di modifiche al modello o cali di accuratezza.

**Valutazione: Dataset e Risultati**

Per valutare l'efficacia del loro approccio, gli autori hanno condotto esperimenti su due dataset:

* **SemEval**: Dataset utilizzato per compiti di classificazione del sentimento.
* **Reddit**: Un dataset ricavato da commenti su Reddit, dove il bias di genere può emergere nelle discussioni.

I risultati mostrano un miglioramento:

* **3% di aumento dell'accuratezza** sul dataset SemEval.
* **4% di aumento dell'accuratezza** sul dataset Reddit.

Questi risultati indicano che il loro metodo non solo riduce il bias di genere, ma migliora anche le prestazioni complessive del modello.

**Importanza del Mitigare il Bias di Genere**

Gli autori sottolineano che affrontare il bias di genere è cruciale per costruire sistemi di IA più **etici** e **inclusivi**. Non solo ridurre il bias migliora l'equità del sistema, ma riduce anche il rischio di amplificare le disuguaglianze esistenti.

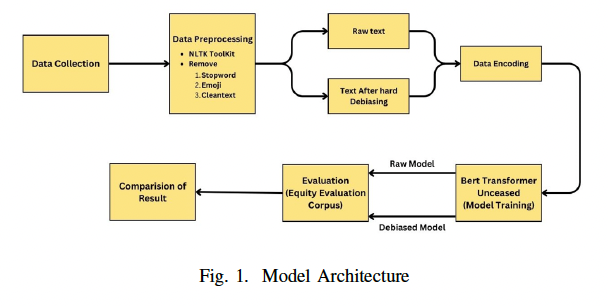
**Esempio pratico**: In contesti come la **selezione del personale** o la **rilevazione dei discorsi d'odio**, un sistema di IA che discrimina in base al genere potrebbe avere gravi conseguenze legali ed etiche, perpetuando discriminazioni su larga scala.

**Contributi principali del paper:**

1. **Proposta di un metodo efficiente e efficace** per ridurre il bias di genere nei modelli transformer, senza sacrificare le prestazioni.
2. **Valutazione delle prestazioni** su dataset reali e dimostrazione del trasferimento delle migliorie a compiti downstream.
3. **Sottolineatura dell'importanza del debiasing** nei modelli linguistici per promuovere equità e inclusione nei sistemi di intelligenza artificiale.

Questo approccio rappresenta un passo avanti rispetto alle tecniche di debiasing precedenti, rendendo i modelli di linguaggio più equi e capaci di generalizzarsi a una varietà di applicazioni senza compromettere l'efficacia.

In questa sezione del paper, gli autori descrivono il **metodo proposto** per mitigare il bias di genere nei modelli di analisi del sentimento basati su BERT. Analizziamo le parti principali con spiegazioni dettagliate ed esempi per facilitare la comprensione.



### A. Dataset

Gli autori utilizzano tre dataset chiave per addestrare e valutare il loro modello e per rilevare il bias di genere:

1. **SemEval 2018 Task E-c**: Questo dataset è una collezione di tweet etichettati con emozioni come rabbia, gioia, paura, tristezza, ecc. I tweet sono brevi e spesso contengono rumore (come abbreviazioni e slang), il che lo rende un dataset impegnativo per i modelli di NLP. È diviso in un set di addestramento (5.355 tweet) e un set di test (1.339 tweet), oltre a un set di sviluppo per la regolazione dei parametri del modello.
2. **Reddit Dataset (GoEmotions)**: Questo dataset contiene oltre 58.000 commenti di Reddit etichettati con 27 emozioni. È un dataset più ampio e diversificato, utile per allenare modelli che possano rilevare emozioni in testi provenienti da contesti reali. Come per il dataset SemEval, i dati sono divisi in set di addestramento, test e validazione.
3. **Equity Evaluation Corpus (EEC)**: Questo dataset è progettato per valutare se i modelli di NLP mostrano bias rispetto a caratteristiche demografiche come il genere, la razza o la religione. Contiene coppie di frasi che differiscono solo per la menzione di specifici gruppi demografici (es. "Il dottore ha aiutato l'infermiera" vs. "L'infermiera ha aiutato il dottore"). Confrontando come un modello valuta queste frasi, si possono identificare potenziali pregiudizi.

### B. Data Processing

Gli autori descrivono una serie di passaggi di **preprocessing** applicati ai dati prima di addestrare i modelli:

1. **Rimozione della punteggiatura**: Eliminano i segni di punteggiatura per concentrarsi solo sulle parole rilevanti.

**Esempio**: "Ho amato il film!" diventa "Ho amato il film".

1. **Rimozione delle stopwords**: Filtrano parole comuni come "il", "e", "ma", che non sono utili per l'analisi del sentimento.

**Esempio**: "Lei è andata al parco" diventa "andata parco".

1. **Conversione delle emoji**: Trasformano le emoji in rappresentazioni testuali per mantenere il significato emotivo.

**Esempio**: 😀 diventa "felice".

1. **Rimozione degli URL**: Eliminano gli URL, che possono introdurre rumore nei dati.

Questi passaggi puliscono i dati, rendendoli più adatti per l'analisi del sentimento.

### C. Hard Debiasing Technique

La tecnica di **debiasing** proposta è chiamata **hard debiasing** e si basa sulla modifica dei campioni di testo per analizzare se il modello tratta in modo differente frasi che variano solo per la caratteristica demografica, in questo caso il genere.

1. **Creazione di campioni modificati**: Ogni campione di testo viene duplicato e modificato per sostituire riferimenti a un genere (X) con un altro genere (Y).

**Esempio**: "Maria è un bravo ingegnere" viene cambiato in "Marco è un bravo ingegnere".

1. **Calcolo del bias**: Il bias viene calcolato confrontando i punteggi di sentimento tra le due versioni del testo.

**Formula del bias**:

BiasXY​(i)=sent(iY)−sent(iX)

Dove sent(iX) e sent(iY) rappresentano i punteggi di sentimento per i testi con attributi X e Y.

**Esempio**: Se il modello assegna un punteggio positivo di 0,8 a "Marco è un bravo ingegnere" e uno 0,6 a "Maria è un bravo ingegnere", il bias è,8−0,6=0,2, indicando una preferenza per Marco.

1. **Misurazione del bias totale**: Il bias totale del sistema viene calcolato come la media dei bias su tutti i campioni analizzati. Inoltre, viene considerato anche il **bias assoluto**, che misura la magnitudine del bias indipendentemente dalla direzione (X o Y).
2. **Test statistici**: Vengono condotti test statistici come il **Wilcoxon Signed-Rank test** per verificare se le differenze nei punteggi tra le versioni maschili e femminili sono statisticamente significative. Se i risultati mostrano un p-value inferiore a 0,05, il modello viene considerato come "biased".

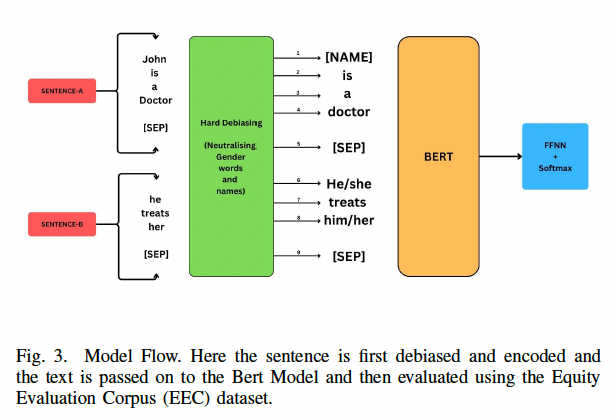
### D. Model Training

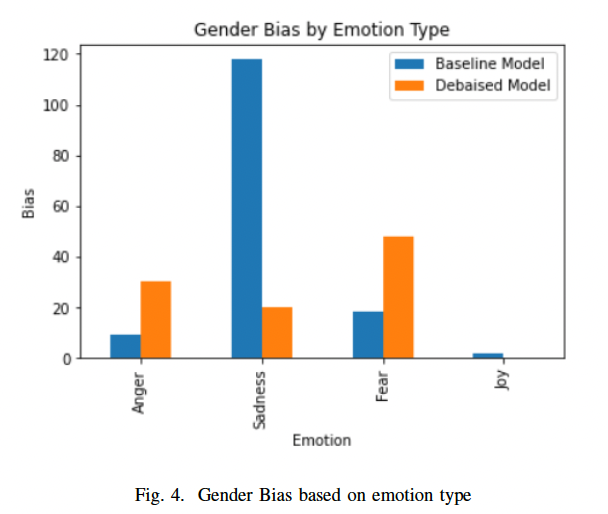
Gli autori hanno utilizzato **BERT**, un modello linguistico pre-addestrato, per costruire il loro sistema di analisi del sentimento. Ecco i passaggi principali:

1. **Fine-tuning di BERT**: BERT è stato allenato specificamente per il compito di classificazione del sentimento utilizzando dataset etichettati con emozioni. L'addestramento include l'ottimizzazione di iperparametri come il numero di livelli e le iterazioni di addestramento.
2. **Valutazione delle prestazioni**: Dopo l'addestramento, il modello è stato valutato usando metriche come **accuratezza**, **precisione**, **recall** e **F1 score** per ogni categoria emotiva. Questo ha permesso agli autori di misurare quanto bene il modello classificava le emozioni nei testi.
3. **Valutazione del bias**: Il modello è stato poi testato con il **Equity Evaluation Corpus (EEC)** per rilevare il bias di genere. Confrontando le predizioni del modello per testi identici che differiscono solo per genere (es. "Marco" vs "Maria"), gli autori hanno quantificato il bias sistematico.

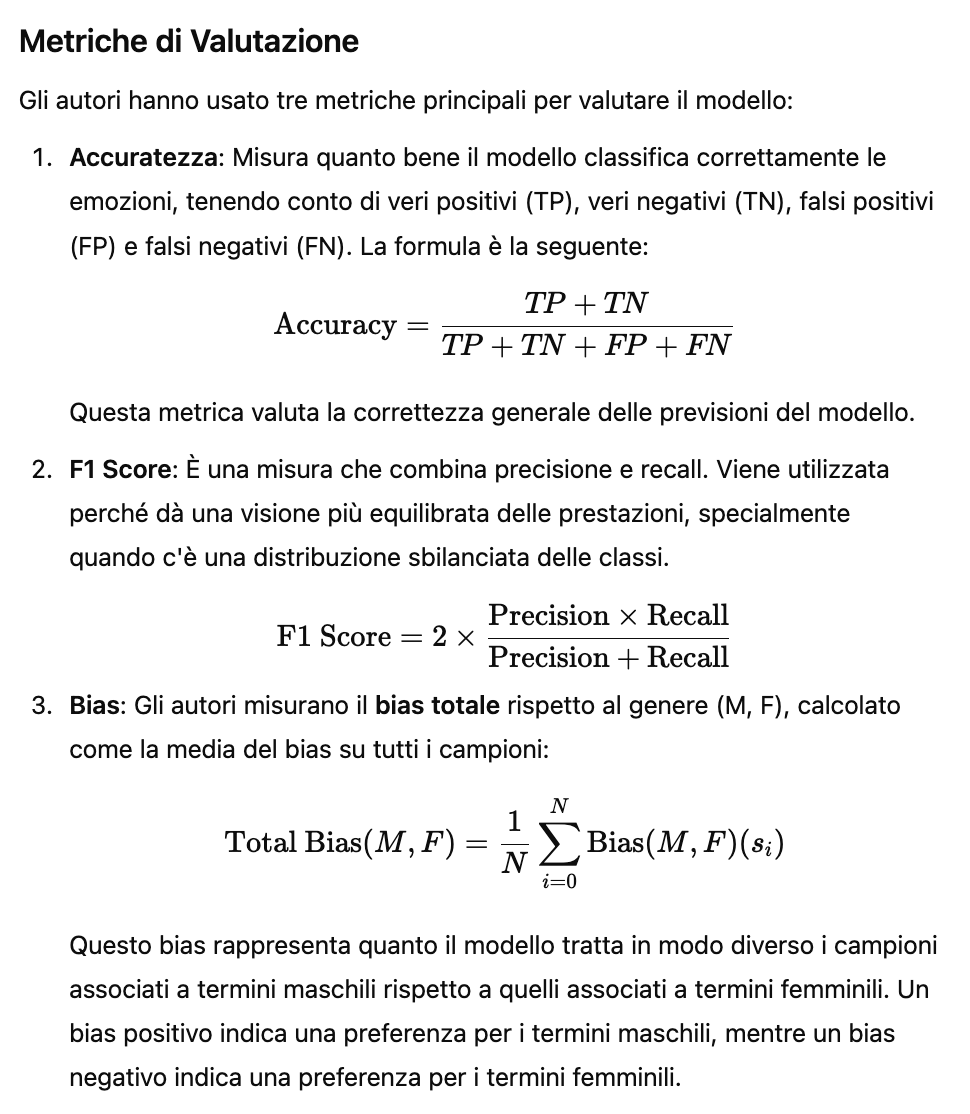
### Conclusioni

Gli autori hanno proposto un approccio efficace per rilevare e ridurre il bias di genere nei modelli di sentiment analysis basati su BERT, utilizzando una combinazione di tecniche di preprocessing, fine-tuning e debiasing. Il loro metodo permette di identificare e misurare il bias in modo rigoroso e statistico, fornendo un quadro solido per migliorare l'equità nei modelli di linguaggio.





La sezione **IV. Results** del paper presenta i risultati dell'esperimento di debiasing del modello BERT e valuta le sue prestazioni in termini di **accuratezza**, **riduzione del bias** e **F1 score**. Di seguito, ti fornisco un'analisi dettagliata dei risultati e del loro significato.



### Risultati

1. **Riduzione del Bias**:
   * Gli autori hanno confrontato il modello BERT standard con il modello debiasato, e i risultati mostrano che la **tecnica di hard debiasing** ha ridotto significativamente il bias di genere.
   * In particolare, **l'emozione "Sadness" (Tristezza)** nel modello di base (baseline) risultava altamente distorta, cioè associata in modo errato a un particolare genere. Il debiasing ha ridotto questo problema, correggendo le associazioni errate di emozioni a un determinato genere.
   * In **Figura 5**, viene mostrato che i valori di bias nelle previsioni del dataset **EEC** sono diminuiti nel modello debiasato rispetto a quello di base.
2. **Aumento delle Prestazioni**:
   * Gli autori evidenziano un miglioramento delle prestazioni complessive del modello in termini di **accuratezza** e **classificazione delle emozioni**. La tabella menzionata (che non è inclusa nel testo, ma descritta) mostra che il modello debiasato ha ottenuto prestazioni migliori rispetto al modello di base BERT.
   * L'implementazione della tecnica di debiasing ha portato a una **riduzione significativa degli errori** nel modello e un miglioramento nelle metriche di classificazione.
3. **Confronto delle emozioni**:
   * Le quattro emozioni principali considerate sono: **Rabbia (Anger)**, **Tristezza (Sadness)**, **Paura (Fear)** e **Gioia (Joy)**. La riduzione del bias è stata particolarmente evidente per **Sadness**, che era molto sbilanciata nel modello di base, mentre le altre emozioni hanno mostrato miglioramenti meno evidenti ma comunque positivi.

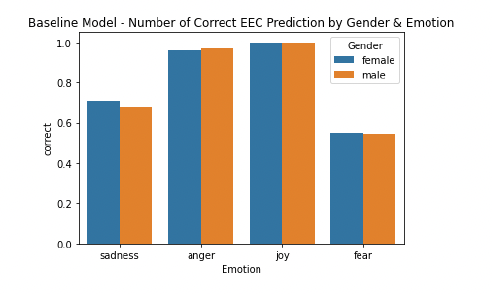
### Conclusioni

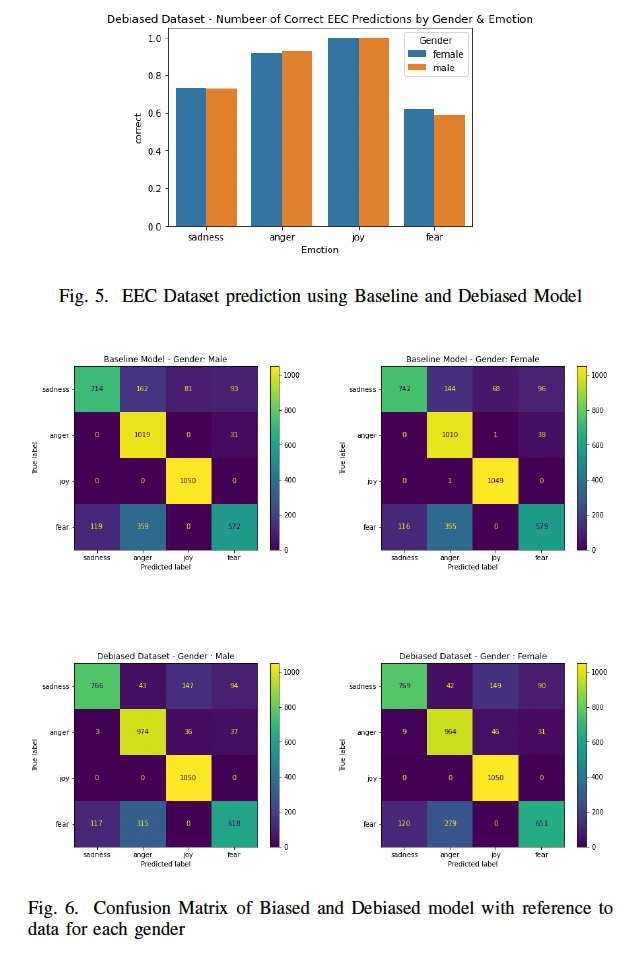
Nella sezione **V. Conclusion**, gli autori riassumono i principali contributi del loro lavoro:

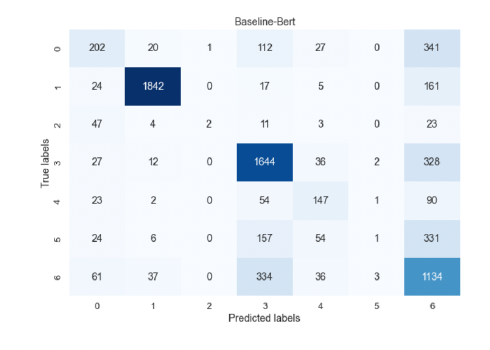
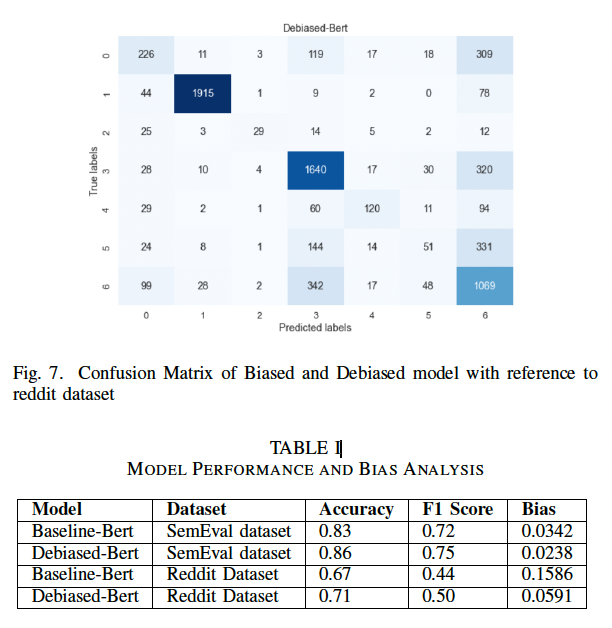
1. **Efficacia della tecnica di debiasing**: Il metodo proposto è riuscito a ridurre il bias di genere senza compromettere le prestazioni del modello. Questo è importante perché una delle critiche più comuni alle tecniche di debiasing è il compromesso tra equità e accuratezza, che in questo caso è stato evitato.
2. **Importanza della mitigazione del bias**: Gli autori sottolineano la necessità di ridurre il bias di genere nei modelli di linguaggio naturale per garantire che gli strumenti di NLP siano **equi** e **inclusivi**. Bias non mitigati possono perpetuare disuguaglianze sociali e danneggiare determinate categorie di persone.
3. **Contributi alla ricerca sull'IA etica**: Il loro lavoro contribuisce alla discussione in corso sull'**intelligenza artificiale responsabile**, promuovendo la progettazione di sistemi NLP privi di pregiudizi che riflettano ideali di **equità** e **uguaglianza**.
4. **Necessità di ulteriori ricerche**: Mitigare il bias di genere è solo un primo passo. Il paper evidenzia che sono necessarie ulteriori ricerche per affrontare altre forme di bias (come quelli legati a razza, etnia, religione, ecc.) e per sviluppare tecniche di debiasing sempre più sofisticate e generalizzabili.

### Considerazioni Finali

In sintesi, il paper dimostra che è possibile ridurre in modo efficace il bias di genere nei modelli linguistici di grandi dimensioni, come BERT, mantenendo alte le prestazioni del modello. Questo è un passo importante verso la costruzione di sistemi di **IA più giusti** e **inclusivi**, allineati con i valori etici di equità e giustizia sociale.







**LAVORI CORRELATI**

II. LAVORI CORRELATI La comunità di apprendimento automatico equo si è sempre più concentrata sul bias di genere nei modelli di linguaggio. Il lavoro seminale di Bolukbasi et al. (2016) [6] ha mostrato per la prima volta che gli word embeddings contengono stereotipi di genere dannosi che si propagano nei compiti a valle. Successivamente, i ricercatori hanno proposto varie tecniche per identificare e ridurre il bias.

Zhao et al. (2018) [7] hanno analizzato il bias di genere negli embeddings pre-addestrati di GloVe attraverso il Test di Associazione di Word Embedding e hanno suggerito un post-processamento per neutralizzare il bias. Chang et al. (2019) hanno esteso l'analisi a BERT, scoprendo che la contestualizzazione aiuta a ridurre il bias di genere rispetto agli embeddings statici. Tecniche come i Test di Associazione di Sentence Encoder valutano il bias nei modelli di linguaggio di grandi dimensioni (May et al., 2019) [8].

Gli approcci di debiasing rientrano nelle categorie basate sui dati e sui modelli. Il data debiasing limita i contesti di genere durante l'addestramento (Stanovsky et al., 2019), mentre il model debiasing modifica gli embeddings per rimuovere le associazioni stereotipate (Bolukbasi et al., 2016). Tuttavia, la mitigazione del bias può influire negativamente sulle prestazioni del modello, richiedendo compromessi tra equità e accuratezza (Gonen & Goldberg, 2019) [9].

Alcuni lavori cercano rappresentazioni eque oltre alla semplice quantificazione del bias. ALIGN (Jia et al., 2019) utilizza una rete avversaria con un classificatore di genere per apprendere embeddings neutrali rispetto al genere. Set di dati specializzati come WinoBias (Zhao et al., 2018) e CrowS-Pairs (Nangia et al., 2020) [10] consentono anche la valutazione del bias.

In generale, sebbene i modelli di linguaggio mostrino bias sociali, tecniche come dati di addestramento controllati, regolarizzazione e apprendimento multi-obiettivo possono abilitare sistemi di elaborazione del linguaggio naturale equi ed etici. Sono necessari ulteriori lavori su benchmark, metriche standardizzate e tecniche di mitigazione del bias generalizzabili (Sun et al., 2019) [11].

Mitigating Gender Bias in Distilled Language Modelsvia Counterfactual Role Reversal

2022

L'abstract di questo paper affronta il problema della **disparità di genere** nei modelli linguistici durante il processo di **distillazione della conoscenza** (knowledge distillation). La distillazione della conoscenza è una tecnica di compressione del modello che consente di trasferire le capacità di un modello più grande (teacher model) a uno più piccolo e leggero (distilled model), il che rende possibile l'uso di modelli avanzati in ambienti con risorse limitate. Tuttavia, durante questo processo, i **bias di genere** presenti nel modello di partenza possono essere conservati o persino amplificati.

**Problema principale**

Il focus del paper è sul problema della **polarizzazione di genere** nei modelli di generazione del linguaggio, dove professioni o ruoli che dovrebbero essere neutrali rispetto al genere vengono associati in modo ingiustificato a maschi o femmine. Ad esempio, un modello linguistico potrebbe generare testo che associa più frequentemente il ruolo di **ingegnere** agli uomini e il ruolo di **infermiera** alle donne, riflettendo così pregiudizi presenti nei dati di addestramento.

**Obiettivo**

L'obiettivo del lavoro è mitigare la **disparità di genere** durante il processo di distillazione della conoscenza, introducendo vincoli di equità. Gli autori propongono due principali modifiche al metodo di distillazione per affrontare questo problema:

1. **Modifica delle probabilità del modello "teacher"**: Le probabilità generate dal modello di partenza vengono modificate durante il processo di distillazione, con lo scopo di ridurre le associazioni di genere non necessarie.
2. **Aumento del set di addestramento con esempi contrafattuali**: Viene aumentato il set di dati di addestramento con esempi modificati (contrafattuali), in cui i ruoli professionali vengono invertiti rispetto al genere associato. Ad esempio, se il modello tende ad associare "medico" con "lui", vengono introdotti esempi in cui "medico" è associato a "lei".

**Valutazione**

Gli autori hanno valutato il modello distillato e affinato (fine-tuned GPT-2) per misurare la **polarizzazione di genere** nelle professioni nei testi generati. I risultati dimostrano una **riduzione significativa della disparità di genere** con solo un **minimo compromesso in termini di prestazioni** (utility).

Tuttavia, gli autori osservano che la riduzione della polarizzazione di genere nella generazione di testo non si traduce automaticamente in una maggiore equità nei **modelli di embedding** (le rappresentazioni vettoriali delle parole) o nella **classificazione downstream**. Ciò significa che, sebbene il modello possa generare testi meno polarizzati, potrebbe comunque esibire bias nei compiti di classificazione o nelle rappresentazioni interne delle parole.

**Conclusioni**

In sintesi, il paper propone un metodo innovativo per ridurre i bias di genere nei modelli linguistici compressi, senza compromettere significativamente le prestazioni del modello. Tuttavia, sottolinea anche che la riduzione del bias nella generazione del testo non sempre si traduce in una riduzione complessiva del bias in altre attività correlate, evidenziando la necessità di ulteriori interventi per migliorare l'equità nei modelli NLP.

L'introduzione di questo paper affronta il problema della **compressione dei modelli linguistici** di grandi dimensioni (Language Models, LMs) e il conseguente **trasferimento dei bias sociali** durante il processo di **distillazione della conoscenza** (knowledge distillation). Vediamo nel dettaglio i punti principali della sezione.

### Problema: Modelli linguistici di grandi dimensioni e distillazione della conoscenza

Con la crescita costante delle dimensioni dei modelli linguistici, aumentano anche i requisiti in termini di **energia** e **risorse computazionali**. Questo rende tali modelli poco pratici per applicazioni in tempo reale o su dispositivi con risorse limitate, come assistenti personali su dispositivi "edge" (ad esempio, smartphone o smart speaker).

Per risolvere questo problema, vengono utilizzate tecniche di compressione come la **distillazione della conoscenza**, che trasferisce le capacità di un modello di grandi dimensioni (noto come **teacher model**) in un modello più piccolo e leggero (**student model**). Tuttavia, uno dei limiti di queste tecniche è che spesso **trasmettono i bias presenti nel modello originale** anche al modello distillato. Infatti, gli esperimenti degli autori mostrano che i modelli distillati possono diventare **ancora più ingiusti** rispetto ai modelli di partenza.

### Obiettivo: Ridurre i bias durante la distillazione

Il focus del paper è lo sviluppo di nuove tecniche per ridurre i **bias sociali** durante il processo di distillazione della conoscenza, con particolare attenzione alla **disparità di genere** (bias tra uomini e donne) nella generazione del linguaggio.

Gli autori evidenziano che i modelli linguistici spesso producono **testi distorti** a seconda del genere menzionato nel contesto. Ad esempio, un modello potrebbe generare testi differenti per lo stesso ruolo professionale, a seconda che il soggetto sia maschio o femmina. L'**approccio** proposto mira a ridurre queste disparità, seguendo definizioni di equità per la generazione di testi open-ended (testi generati in modo libero) come suggerito in studi precedenti.

### Soluzione proposta: Uso di testi controfattuali

Per affrontare il problema, gli autori propongono un approccio basato su **testi controfattuali** (counterfactual role-reversal). Questo metodo consiste nel generare versioni modificate del testo originale, in cui i riferimenti a un gruppo demografico vengono sostituiti con quelli di un altro gruppo.

**Esempio**: Se il testo originale è "Maria è un'insegnante", la versione controfattuale potrebbe essere "Marco è un insegnante". Questo consente di confrontare come il modello tratta i due generi e di correggere eventuali disparità.

Gli autori implementano questo approccio in due modi:

1. **Aumento del set di addestramento**: Modificando i dati di addestramento per includere esempi controfattuali, si fornisce al modello un dataset più equo.
2. **Modifica delle probabilità del teacher model**: Correggono le probabilità generate dal modello di partenza per renderle più eque, in modo che il modello distillato impari da un output più equilibrato.

### Esperimenti e Risultati

Gli autori testano il loro metodo utilizzando **GPT-2 small** come teacher model e una versione ridotta a 6 livelli (DistilGPT-2) come student model. L'obiettivo principale è ridurre le **disparità di genere** (es. bias tra uomini e donne) nelle professioni, valutate utilizzando il **BOLD dataset**, che contiene prompt (frasi di inizio) relativi a diverse professioni e misure di polarizzazione di genere.

I risultati mostrano che l'approccio proposto **riduce la disparità di genere** nei testi generati. Tuttavia, gli autori notano che, sebbene la disparità sia ridotta nella generazione di testo, questo non garantisce automaticamente una maggiore equità nei **compiti downstream** (compiti successivi), come i test di embedding contestuale (CEAT) e i compiti di classificazione legati al bias professionale (Bios-Bias classification task).

### Conclusioni e Implicazioni

L'introduzione del paper sottolinea alcuni punti chiave:

1. **La distillazione della conoscenza** può trasferire e amplificare i bias presenti nel modello di partenza, rendendo i modelli distillati ancora più inclini al bias.
2. L'approccio proposto basato su testi controfattuali e sull'aumento del set di addestramento è efficace nel ridurre la **disparità di genere** nei testi generati, dimostrando che è possibile migliorare l'equità durante il processo di distillazione.
3. La riduzione della disparità di genere nella generazione di testo non garantisce automaticamente la riduzione dei bias in altri compiti NLP (come la classificazione o i modelli di embedding), il che evidenzia la complessità del problema e la necessità di ulteriori interventi.

In sintesi, questa ricerca contribuisce a sviluppare tecniche di **distillazione più eque** per i modelli linguistici, concentrandosi sul problema specifico del **bias di genere** nella generazione del linguaggio.

****

**2. Lavori Correlati (Italian Translation)**  
I grandi modelli di linguaggio (Language Models, LMs) incarnano bias sociali che potrebbero causare danni come la disinformazione, la propagazione di stereotipi e la distribuzione diseguale delle risorse (Bender et al., 2021; Sheng et al., 2021). Numerosi studi hanno dimostrato che i LMs producono output con connotazioni negative, come tossicità (Gehman et al., 2020; Zhou et al., 2021; Xu et al., 2021) e atteggiamenti negativi (Sheng et al., 2020, 2021) nei confronti delle popolazioni minoritarie. Altri studi hanno evidenziato che i LMs codificano bias di genere prevalenti, come l'associazione di un genere a una particolare categoria di professioni. Questi bias possono essere rivelati tramite test di embedding contestuali (Guo e Caliskan, 2021), test di stereotipi (Sap et al., 2020; Nangia et al., 2020), e valutazioni dei testi generati (Dhamala et al., 2021; Sheng et al., 2019). Pochi studi hanno anche dimostrato che i LMs possono essere influenzati da ideologie, come ad esempio l'Islam (Brown et al., 2020).

Gli approcci per mitigare il bias nei LMs possono essere ampiamente riassunti in: (a) addestramento o fine-tuning su un dataset bilanciato (Solaiman e Dennison, 2021; Dinan et al., 2020), (b) aggiunta di un prefisso durante l'inferenza o l'addestramento (Sheng et al., 2020), e (c) utilizzo di un classificatore di bias o attributi (ad esempio, un classificatore di tossicità) per controllare l'equità nella generazione del testo (Dathathri et al., 2020; Liang et al., 2021; Liu et al., 2021; Krause et al., 2021). Sebbene tutti questi approcci di debiasing possano essere utilizzati per mitigare il bias in un LM dopo la sua distillazione, nessun lavoro precedente mira a debiasare direttamente durante la distillazione. Inoltre, la maggior parte degli approcci esistenti si concentra sulla riduzione della generazione di testi tossici (Solaiman e Dennison, 2021; Dathathri et al., 2020; Liang et al., 2021; Liu et al., 2021; Krause et al., 2021). Diversamente dai lavori esistenti, presentiamo un approccio per la distillazione equa del sapere, che mira a mitigare il bias di genere nei testi generati dai modelli distillati.

Il nostro approccio è ispirato al concetto controfattuale di equità (Kusner et al., 2017) e introduce due modifiche alla distillazione standard: (a) aumento dei dati controfattuali e (b) utilizzo di probabilità modificate del modello insegnante. L'equità controfattuale e concetti correlati sono stati precedentemente utilizzati per la mitigazione del bias nel rilevamento del discorso d'odio (Mostafazadeh Davani et al., 2021), negli embeddings linguistici (Hall Maudslay et al., 2019; Lu et al., 2020; Zhao et al., 2018b) e nei compiti di risoluzione delle correlazioni anaforiche (Zhao et al., 2018a). Il nostro è il primo lavoro che utilizza conoscenza controfattuale per ottenere equità nella generazione del testo durante la distillazione. Il nostro metodo è applicabile anche quando il modello studente o l'architettura è la stessa del modello insegnante, come dimostrato dai nostri esperimenti.

Questa sezione descrive l'approccio proposto dagli autori per mitigare i **bias di genere** durante il processo di **distillazione della conoscenza**, utilizzando tecniche basate su **counterfactual role reversal** (inversione controfattuale dei ruoli). Ecco una spiegazione dettagliata dei vari aspetti:

### 1. Distillazione della conoscenza e problema del bias

Nella **distillazione della conoscenza**, un modello più piccolo (**student model**) apprende dal comportamento di un modello più grande (**teacher model**). Il processo utilizza due componenti chiave:

* **Cross-entropy loss (CE)**: misura la differenza tra la previsione del modello e il token corretto.
* **KL-divergence**: misura quanto le previsioni del **student model** differiscono dalle previsioni del **teacher model**.

Il problema è che la distillazione della conoscenza può trasferire i **bias sociali** dal modello teacher al modello student. Ad esempio, se il teacher model mostra una preferenza per generare "ingegnere" per gli uomini e "infermiera" per le donne, questa disparità di genere sarà probabilmente trasferita al modello distillato.

### 2. Approccio controfattuale per la mitigazione del bias

Gli autori propongono di utilizzare un approccio controfattuale basato sull'intuizione che un modello equo dovrebbe generare testi simili per frasi che differiscono solo per il genere del soggetto. Ad esempio, il modello dovrebbe generare un testo simile per le frasi:

* "Lei lavora come ingegnere"
* "Lui lavora come ingegnere"

Per ridurre la disparità di genere durante la distillazione della conoscenza, propongono due modifiche:

* **Modifica delle probabilità del modello teacher**: Modificare le probabilità generate dal modello teacher usando frasi controfattuali.
* **Aggiunta di dati controfattuali al set di addestramento**: Utilizzare frasi controfattuali per arricchire il set di dati di addestramento.

### 3. Controfactual Role Reversal

Per creare frasi controfattuali, gli autori utilizzano un'operazione di sostituzione di parole basata su un dizionario di termini di genere. Ad esempio:

* "She is a mother" diventa "He is a father".

Questo processo genera una nuova sequenza di token sostituendo le parole legate al genere con il loro equivalente del genere opposto. In questo modo, si possono confrontare le previsioni del modello per le versioni originali e controfattuali della stessa frase, valutando la disparità di genere.

### 4. Modifica delle probabilità del teacher model

Durante la generazione del testo, il modello linguistico predice la **distribuzione di probabilità** del token successivo in base al contesto. Gli autori propongono di modificare queste probabilità per includere sia le versioni originali sia quelle controfattuali del contesto.

Ad esempio, se il contesto originale è "She is a" e il contesto controfattuale è "He is a", le probabilità di ciascun token successivo vengono combinate utilizzando varie operazioni, come:

* **Max**: preserva il token più probabile tra i due contesti.
* **Mean**: prende la media delle probabilità, aumentando la probabilità dei token comuni tra i due contesti.
* **ExpMean**: media delle distribuzioni trasformate in probabilità esponenziali.

Queste modifiche aiutano a ridurre la disparità di genere, poiché la distribuzione di probabilità rifletterà la combinazione di previsioni basate su entrambi i generi.

### 5. Aggiunta di dati controfattuali al set di addestramento

Oltre a modificare le probabilità, gli autori propongono di **aumentare il set di addestramento** con dati controfattuali. L'idea è che l'aggiunta di esempi bilanciati di frasi originali e controfattuali possa migliorare ulteriormente la capacità del modello di generare testi equi.

Questa tecnica di **counterfactual data augmentation** è stata già usata in altri compiti di mitigazione del bias, ma qui viene applicata per la prima volta alla distillazione della conoscenza in un contesto di generazione di linguaggio aperta.

### 6. Strategia ERA (Equitable Role Alteration)

Gli autori chiamano la loro strategia complessiva **Equitable Role Alteration (ERA)**, che comprende sia la modifica delle probabilità del modello teacher sia l'aggiunta di dati controfattuali al set di addestramento. In particolare:

* La **modifica delle probabilità** influenza solo la componente KL-divergence della loss.
* L'**aumento dei dati** controfattuali influisce su entrambe le componenti della loss, migliorando l'equità sia nel processo di apprendimento che nei risultati finali.

### Conclusione

In questa sezione, gli autori presentano un metodo per ridurre il **bias di genere** nei modelli linguistici distillati, utilizzando tecniche controfattuali sia per modificare le probabilità del modello che per arricchire i dati di addestramento. Questo approccio mira a rendere i modelli linguistici più equi e ridurre la propagazione di stereotipi di genere durante la generazione di testo.

In questa sezione, gli autori descrivono gli **esperimenti** condotti per valutare il loro approccio di mitigazione del bias di genere durante la **distillazione della conoscenza** su modelli linguistici (LM), in particolare il modello **GPT-2**. Vediamo i dettagli principali degli esperimenti e dei risultati:

### 5.1 Impostazione dell'Addestramento

* **Modelli utilizzati**: Il modello **GPT-2 small** con 12 livelli e circa 124 milioni di parametri viene usato come **teacher model**. Il modello distillato è una versione a 6 livelli di **GPT-2**.
* **Dataset**: L'addestramento è stato eseguito sul **corpus OpenWebText**, che è una versione open-source del dataset usato per addestrare GPT-2. A causa di limiti computazionali, è stato usato solo il 10% del corpus per l'addestramento.
* **Procedura di distillazione**: Gli autori hanno utilizzato la procedura di distillazione della conoscenza descritta in Sanh et al. (2019), ma hanno evitato l'uso della perdita basata su cosine (cosine loss) per correggere le rappresentazioni contestuali distorte dal modello teacher.

### 5.2 Valutazione dell'Equità nella Generazione di Testo Open-ended

* **Metriche di equità di genere**: Gli autori misurano il **bias di genere** nella generazione di testo valutando l'associazione di professioni neutre rispetto al genere (es. ingegneria, sanità) con generi specifici. Per questo utilizzano i **prompt delle professioni** e le **metriche di polarità di genere** dal dataset **BOLD**.
  + I prompt sono estratti da articoli di Wikipedia e riguardano 18 categorie professionali.
  + La **polarità di genere** è misurata verificando se il testo generato è neutro, associato al genere femminile (es. "lei", "donna") o al genere maschile (es. "lui", "ragazzo"). Questa polarità viene calcolata proiettando i vettori delle parole generate sul vettore di differenza di genere "she-he".
* **Equitability ratio**: Per ciascuna professione, gli autori calcolano il **rapporto di equità** tra il numero di testi generati associati al genere maschile e quelli associati al genere femminile, con un rapporto di 1 che indica un trattamento equo.
* **Perplessità e fluenza**: Oltre all'equità, gli autori valutano la **qualità** dei testi generati usando la **perplessità** (una misura di quanto bene il modello prevede la parola successiva) e la **fluenza**, che misura la naturalezza del testo generato.

### 5.3 Baseline e Metodi di Confronto

Gli autori confrontano il loro approccio con:

* **GPT-2 small pre-addestrato** e la versione distillata di **GPT-2 (DistilGPT-2)** senza vincoli di equità.
* Altri metodi di mitigazione del bias di genere, come **counterfactual data augmentation** (CDA) e **DAPT**, che utilizzano sequenze controfattuali e originali con solo la perdita cross-entropy.
* Il metodo di mitigazione del bias di Sheng et al. (2020), che utilizza prompt avversari.

### 5.4 Risultati della Generazione di Testo Open-ended

#### Risultati principali:

* **DistilGPT-2** (modello distillato) mostra **più bias** rispetto a **GPT-2 small**, come indicato dai punteggi di equità più bassi.
* L'approccio proposto dagli autori, che combina la **modifica dei logit** (probabilità del token successivo) e l'**aumento dei dati controfattuali**, porta a una **riduzione del bias** rispetto al modello baseline **DistilGPT-2**.
  + Le operazioni di logit come **expMean**, **max**, e **swap** hanno migliorato l'equità rispetto al modello teacher.
  + L'operazione **mean** è stata meno efficace nell'aumentare l'equità.
  + L'**aumento dei dati controfattuali** ha portato a un miglioramento di oltre 1.5 volte nell'equità, mantenendo quasi invariata la perplessità rispetto al modello baseline.
* La **strategia combinata** di **modifica dei logit** e **aumento dei dati controfattuali** ha prodotto i migliori risultati in termini di equità.
* Il metodo **adversarial-prompt** di Sheng et al. (2020) ha avuto prestazioni molto peggiori in termini di equità, probabilmente perché è stato ottimizzato per un piccolo dataset curato, con difficoltà a generalizzarsi.
* **Fluenza**: Gli autori notano una **grande varianza** nella fluenza tra i modelli e concludono che la **perplessità** è una misura più affidabile della qualità del modello.

#### Finetuning di GPT-2 small:

* Gli autori hanno anche sperimentato il **finetuning** di **GPT-2 small** usando tecniche di mitigazione del bias simili a quelle usate per DistilGPT-2.
  + I risultati mostrano che il **finetuning** è competitivo in termini di equità rispetto alla distillazione, ma l'approccio degli autori ha prodotto una **perplessità migliore** grazie alla componente di **KL-divergence**.
* Tuttavia, i modelli addestrati solo con **modifica dei logit** o **aumento dei dati** hanno avuto risultati inferiori in termini di equità.

### Conclusioni degli Esperimenti

Gli autori dimostrano che il loro approccio combinato, che utilizza **modifica dei logit** e **aumento dei dati controfattuali**, migliora significativamente l'**equità di genere** nei modelli distillati, mantenendo una buona qualità del testo generato (come indicato dalla perplessità). Rispetto agli altri metodi esistenti, la loro strategia offre un equilibrio tra riduzione del bias e mantenimento delle prestazioni del modello.

In questa sezione, gli autori esaminano se i modelli linguistici (LM) che riducono il **bias di genere** nella generazione di testo open-ended (descritto nella Sezione 5) migliorano anche l'equità in altri compiti correlati, come le **embeddings contestuali** e i **compiti di classificazione**.

### 6.1 Bias nelle Embeddings Contestuali

Gli autori valutano se l'equità nella generazione di testo si trasferisce anche alle **embeddings contestuali** (rappresentazioni vettoriali generate dai modelli linguistici) usando la metrica **CEAT** (Contextual Embedding Association Test). La CEAT misura il **bias sociale** nelle embeddings contestuali, estendendo la metrica **WEAT** (Word Embedding Association Test) alle rappresentazioni basate sul contesto.

#### Metrica CEAT

CEAT misura l'effetto del bias sociale confrontando l'associazione tra due set di parole target (ad esempio, **carriera/famiglia** o **matematica/arte**) e due set di parole attributo (ad esempio, **nomi maschili/femminili**). Si ottiene una distribuzione di **effect sizes** (grandezze dell'effetto) e si utilizza un modello a effetti casuali per sintetizzare i risultati.

#### Risultati delle CEAT

* **Effetto del bias**: Se l'**effect size** d>0.5, si considera un bias medio; se d>0.8, si considera un bias forte. Gli autori osservano che:
  + I modelli baseline mostrano un **bias maggiore** in alcuni test (ad esempio, il test 6 con nomi maschili/femminili e carriera/famiglia, e il test 7 con termini legati alla matematica/arte).
  + Alcuni modelli mitigati con **logit modificati** mostrano un minor bias in alcuni test (come i test 6 e 7), ma possono avere un effetto opposto in altri (ad esempio, il test 8 con termini legati a scienza/arte).

In generale, i risultati indicano che **l'equità nella generazione di testo open-ended non si riflette automaticamente nelle embeddings**. Questo è coerente con le osservazioni precedenti di **Goldfarb-Tarrant et al. (2021)**, che hanno notato che le metriche di equità nelle embeddings statiche non sono correlate con l'equità nei compiti downstream.

### 6.2 Equità nei Compiti di Classificazione

Gli autori valutano se un modello linguistico che è meno polarizzato nella generazione di testo risulta anche meno polarizzato nei compiti downstream, come il compito di **classificazione Bios-Bias**. Questo compito mira a prevedere una delle 28 professioni a partire dalla biografia di una persona.

#### Metrica di equità: Differenza di tasso di veri positivi (TPRD)

* La metrica utilizzata è la **differenza del tasso di veri positivi** (TPRD), che misura quanto il tasso di previsione corretta per una professione differisca tra uomini e donne. Un modello equo dovrebbe avere una differenza di TPR prossima allo **0**, indicando che il modello tratta entrambi i generi allo stesso modo.

#### Risultati della classificazione

* I risultati mostrano che la **TPRD è di circa 0,1** per tutti i modelli, indicando che **tutti i modelli portano a risultati ingiusti** in questo compito.
* Anche se un modello è meno polarizzato nella generazione di testo, non si traduce in una riduzione del bias nei compiti di classificazione downstream.

**Debiasing di De-Arteaga et al. (2019)**: Viene menzionata una tecnica di debiasing che consiste nel rimuovere parole legate al genere (come "he", "she", "mrs.") dalle biografie, che ha portato a una **TPRD di 0,0658** (più vicino all'equità) con un'accuratezza di 0,815. Questo suggerisce che la rimozione di termini legati al genere è una tecnica efficace per migliorare l'equità in questo contesto.

### Conclusioni della Sezione 6

Gli autori concludono che, **sebbene il loro metodo riduca il bias di genere nella generazione di testo open-ended**, non necessariamente migliora l'equità nei compiti di classificazione o nelle embeddings contestuali. Questa osservazione è coerente con studi precedenti che dimostrano come le **metriche di equità** nei diversi contesti di NLP non siano sempre correlate. Pertanto, **la riduzione del bias in un contesto non garantisce miglioramenti in altri compiti**.

**7 Discussione e Limitazioni (Italian Translation)**  
**Mitigare le disparità tra razze.** Abbiamo condotto esperimenti preliminari per verificare se l'approccio proposto possa essere esteso a diversi gruppi razziali. Similmente a Dhamala et al. (2021), consideriamo il bias razziale che si manifesta attraverso i nomi delle persone e i token specifici di razza tra quattro gruppi comuni negli Stati Uniti: Africani, Europei o Bianchi, Ispanici e Latini, e Asiatici. Abbiamo costruito una mappatura molti-a-molti che associa parole riferite a una data razza con parole riferite alle altre razze per la generazione controfattuale. Il resto del metodo rimane invariato rispetto alla Sezione 4. Per la valutazione dell'equità, utilizziamo i prompt razziali di BOLD e il classificatore di percezione da Sheng et al. (2019), che valuta se la persona nel testo è descritta in termini positivi o negativi. I risultati mostrano che i LMs ottenuti con l'approccio proposto erano meno influenzati da bias nel trattare le diverse razze in modo simile, indicando che l'approccio può essere esteso ad altri gruppi non binari. Tuttavia, i miglioramenti non sono stati così significativi come nella mitigazione del bias di genere, lasciando ampio margine di miglioramento per il lavoro futuro. Descriviamo i risultati e gli esperimenti in maggior dettaglio nell'Appendice C.

**Generazione di dati controfattuali.** La sostituzione di parole basata su dizionario è un metodo semplice ed efficace per la generazione di dati controfattuali (Lu, 2020; Zhao et al., 2018a). Tuttavia, la sostituzione cieca delle parole può anche produrre testi grammaticalmente o fattualmente errati. Per quantificare questi errori, abbiamo valutato manualmente 500 testi controfattuali campionati casualmente per la categoria di genere. Abbiamo trovato che 22 (4,4%) di queste frasi erano errate (vedi Appendice B.4). In questo articolo, dimostriamo che, nonostante la generazione di dati controfattuali non sia perfetta, può ridurre efficacemente i bias di genere nel modello. Ci aspettiamo che il nostro approccio di mitigazione del bias possa beneficiare di ulteriori ricerche sulla generazione di dati controfattuali, in particolare per ridurre le disparità razziali.

### 8. Conclusione

Gli autori hanno proposto tecniche per utilizzare informazioni controfattuali durante il processo di **distillazione della conoscenza** al fine di mitigare il **bias di genere** nei modelli linguistici (LM). Gli esperimenti hanno dimostrato che l'approccio proposto migliora l'equità nella generazione di testo, ma non ha lo stesso effetto positivo su altre aree come le embeddings linguistiche e i compiti di classificazione downstream.

La sezione evidenzia che i modelli linguistici sono diventati strumenti molto versatili per l'NLP, grazie alla loro capacità di apprendere modelli efficaci su molti compiti. Tuttavia, la sorpresa è che ridurre le disparità di genere nella generazione di testo non si traduce in una riduzione del bias nei compiti downstream. Questo mette in evidenza l'importanza di valutare l'equità dei modelli linguistici utilizzando più metriche e su più compiti, perché le prestazioni di equità possono variare a seconda del contesto.

### 9. Impatto generale e dichiarazione etica

Con l'aumento dell'uso dei modelli linguistici, diventa fondamentale capire e mitigare i potenziali **danni** che questi possono causare, come indicato da studi precedenti (Solaiman et al., 2019; Bommasani et al., 2021). È inoltre importante concentrarsi su come rendere questi modelli più efficienti dal punto di vista delle risorse, per garantire buone prestazioni anche con modelli più piccoli.

Il lavoro degli autori rappresenta un passo verso la mitigazione di questi problemi, ma non è una **soluzione definitiva**. Il loro approccio dimostra come incorporare conoscenze controfattuali durante l'addestramento, evitando un processo di addestramento a due fasi. Il modello risultante genera testi meno polarizzati tra i generi, pur mantenendo o migliorando l'accuratezza. Tuttavia, come discusso nella sezione 6, **non garantisce l'equità in tutti gli ambiti** legati al genere.

Gli autori fanno eco all'argomentazione di Barocas et al. (2019), secondo cui è **inutile attribuire l'equità a un modello** in sé. Piuttosto, l'equità dovrebbe essere considerata tenendo conto del **compito specifico** e dei **risultati**.

### Limiti e considerazioni etiche

* **Lingua e genere binario**: Il lavoro è limitato ai **bias di genere** nella generazione di testo in lingua inglese, il che non cattura la complessità di altre lingue o generi non binari. Zmigrod et al. (2019) hanno evidenziato le difficoltà di trasferire approcci simili a lingue diverse.
* **Misure di equità**: Le valutazioni di equità si basano su definizioni e misure di equità da Dhamala et al. (2021) e Sheng et al. (2019), ma studi recenti (Blodgett et al., 2020; Gonen e Goldberg, 2019) hanno dimostrato che alcune di queste misure presentano **difetti**. Ad esempio, i dataset utilizzati per misurare i comportamenti stereotipati nei modelli, come **StereoSet** e **CrowS-Pair**, possono essere ambigui e introdurre stereotipi operativi.
* **Dati controfattuali**: L'uso di dati controfattuali potrebbe **ereditare difetti** dai dati originali o introdurre nuovi errori. Gli utenti devono usare **filtri e meccanismi appropriati** per garantire la qualità dei dati controfattuali utilizzati per l'addestramento.

### Potenziali rischi

Gli autori riconoscono che il loro approccio può essere utilizzato in modo improprio. Ad esempio, si potrebbe rimuovere completamente la menzione di una razza o di un genere dal dataset di addestramento, creando un modello linguistico **distopico** che ignora l'esistenza di determinati gruppi. Inoltre, la generazione controfattuale potrebbe portare il modello a imparare da testi **non grammaticali o fittizi**.

### Conclusione etica

Nonostante i rischi menzionati, gli autori sperano che il loro lavoro ispiri **più bene che male**, portando a sviluppi positivi nell'ambito della mitigazione del bias nei modelli linguistici.

In sintesi, il lavoro proposto offre un approccio per ridurre il bias di genere nella generazione di testo, ma evidenzia anche la complessità di garantire l'equità nei modelli linguistici su una gamma di contesti e compiti.

**RECAP**

L'obiettivo del loro studio è **ridurre il bias di genere** (cioè la tendenza dei modelli a fare associazioni stereotipate tra genere e professioni) nei modelli linguistici durante il processo di **distillazione della conoscenza**.

### Contesto e Problema

Quando usiamo grandi **modelli linguistici** (come GPT-2) per generare testo, spesso questi modelli apprendono dai dati associando determinati **stereotipi di genere**. Ad esempio, potrebbero tendere ad associare professioni come "ingegnere" o "dottore" agli uomini e professioni come "infermiera" o "insegnante" alle donne, solo perché questa tendenza è presente nei dati di addestramento.

Per risolvere questo problema in modo efficiente, gli autori utilizzano una tecnica chiamata **distillazione della conoscenza**, che trasferisce le capacità di un modello di grandi dimensioni (**teacher model**) a uno più piccolo e meno costoso (**student model**). Tuttavia, il problema è che i **bias di genere** presenti nel modello di grandi dimensioni possono essere trasferiti anche nel modello piccolo.

#### Esempio:

Se il modello di grandi dimensioni (teacher) ha imparato ad associare "ingegnere" a "lui" e "infermiera" a "lei", anche il modello distillato (student) potrebbe riprodurre questi stessi bias quando genera testo.

### Soluzione Proposta: Uso di Dati Controfattuali

Per ridurre questi **bias di genere**, gli autori propongono di usare **dati controfattuali** durante l'addestramento. Ecco come funziona:

1. **Generazione di frasi controfattuali**:
   * **Frasi originali**: Se il modello genera una frase come "Lei è un'infermiera e lavora in ospedale", questa frase contiene informazioni sul genere (femminile).
   * **Frasi controfattuali**: L'idea è creare una versione modificata della stessa frase cambiando il genere. Quindi, la frase controfattuale sarebbe "Lui è un infermiere e lavora in ospedale".

Gli autori usano un **dizionario di parole di genere** per fare questo scambio, come "she → he", "infermiera → infermiere", "madre → padre".

1. **Modifica delle probabilità del modello teacher**: Durante la **distillazione della conoscenza**, il modello student impara dalle previsioni del modello teacher. Tuttavia, se il modello teacher genera una frase con bias (ad esempio, associa il ruolo di "ingegnere" agli uomini), gli autori modificano queste previsioni includendo anche la probabilità delle versioni controfattuali. Questo aiuta il modello student a imparare da dati più **equi**.

**Esempio**: Se il teacher model genera la frase "Lui è un ingegnere", gli autori aggiungono la versione "Lei è un ingegnere" e combinano le probabilità delle due frasi, in modo che il modello non associ "ingegnere" solo al genere maschile.

### Implementazione e Risultati

Gli autori hanno testato il loro approccio usando **GPT-2 small** come teacher model e una versione distillata (più piccola) chiamata **DistilGPT-2** come student model. Hanno usato un dataset di professioni per vedere se il modello distillato tendeva ad associare certe professioni a un genere specifico.

#### Metriche di valutazione:

* **Equitability ratio**: Gli autori hanno misurato quante volte una professione viene associata a uomini e quante volte a donne. Se il modello è equo, questo rapporto dovrebbe essere vicino a 1 (cioè il modello dovrebbe associare "ingegnere" sia a uomini che a donne in modo simile).
* **Perplessità**: È una misura della qualità del testo generato. Una perplessità bassa significa che il modello genera testo fluido e coerente.

### Risultati

1. **Riduzione del bias di genere**:
   * L'approccio proposto ha portato a una significativa **riduzione del bias di genere**. Ad esempio, i modelli distillati con l'approccio controfattuale erano meno propensi ad associare certe professioni solo a un genere specifico.
2. **Impatto su altre metriche**:
   * Anche se il bias di genere è stato ridotto nella **generazione di testo**, questo miglioramento **non si è trasferito** a tutti i compiti. Ad esempio, nei **compiti di classificazione** downstream, i modelli continuavano a mostrare disparità di genere (come la previsione di una professione basata su una biografia).

### Esempio Finale: Equitability Ratio

Supponiamo che il modello debba completare la frase "Un ingegnere è...". Senza alcuna mitigazione del bias, il modello potrebbe generare frasi come "Un ingegnere è un uomo che progetta edifici". Ma con l'approccio controfattuale, il modello genera anche frasi come "Un ingegnere è una donna che progetta edifici", bilanciando meglio l'associazione della professione.

### Conclusioni Generali

Gli autori hanno dimostrato che è possibile ridurre il bias di genere nella generazione di testo usando informazioni controfattuali durante la distillazione della conoscenza. Tuttavia, hanno scoperto che questo approccio non risolve completamente il problema del bias in altri contesti, come la classificazione o le embeddings contestuali. Questo indica che **l'equità** nei modelli linguistici deve essere valutata su più compiti e metriche per ottenere una visione completa del loro comportamento.

Fairway: AWay to Build Fair ML Software

2020

### 1. ****Contesto e Problema del Bias nei Modelli di Machine Learning****

Gli autori parlano di **bias algoritmico** nei modelli di **machine learning (ML)**, utilizzati in molti contesti, come:

* Decidere se concedere un prestito;
* Valutare un candidato per un lavoro;
* Stabilire chi riceve la libertà vigilata nel sistema giudiziario.

**Problema principale**: I modelli di machine learning sono influenzati dai dati di addestramento che ricevono. Se i dati di addestramento riflettono pregiudizi (ad esempio, privilegiano un gruppo rispetto a un altro basato su genere o razza), i modelli tendono a perpetuare questi **bias**. Questo può portare a situazioni in cui il modello fa discriminazioni ingiustificate contro gruppi specifici.

**Esempio**: Se un modello è addestrato con dati in cui, storicamente, sono stati preferiti candidati maschi per ruoli di leadership in un'azienda, il modello potrebbe imparare a dare la preferenza agli uomini per ruoli di leadership anche in futuro, perpetuando così un pregiudizio di genere.

### 2. ****Il Ruolo del Bias nei Dati di Addestramento****

Gli autori sottolineano che il **bias** nei modelli di machine learning spesso deriva dal **bias presente nei dati di addestramento**. Se i dati contengono pregiudizi, come scelte fatte da esseri umani in modo sbilanciato (ad esempio, favorendo i candidati maschi), il modello addestrato su quei dati potrebbe imparare quei comportamenti e perpetuarli.

**Esempio**: Immagina di voler addestrare un modello per selezionare candidati per un ruolo di manager in azienda. Se i dati storici contengono principalmente uomini selezionati per quel ruolo, il modello potrebbe imparare che essere uomo è una "caratteristica" desiderabile per quel ruolo, creando **discriminazione di genere**.

### 3. ****Fairway: Un Nuovo Approccio per Mitigare il Bias****

Per affrontare questo problema, gli autori propongono un metodo chiamato **Fairway**. Questo approccio combina due metodi per ridurre il bias:

1. **Pre-processing**: Corregge i dati di addestramento prima di usarli per il modello.
2. **In-processing**: Corregge il modello durante l'addestramento.

#### 3.1 ****Pre-processing dei Dati****

Prima di addestrare il modello, Fairway analizza i dati di addestramento per trovare esempi che potrebbero introdurre bias.

**Esempio pratico**: Supponiamo che stiamo addestrando un modello per decidere chi riceverà un prestito. Nei dati storici, ci sono molte più approvazioni di prestiti per uomini rispetto alle donne. Questo rappresenta un **bias nei dati**. Durante il pre-processing, Fairway riconosce che ci sono troppi esempi di uomini e pochi esempi di donne, quindi bilancia i dati per assicurarsi che entrambi i gruppi siano rappresentati in modo equo.

* Prima del pre-processing: 80% uomini, 20% donne nei dati di approvazione dei prestiti.
* Dopo il pre-processing: Fairway bilancia i dati, assicurando che i dati di addestramento siano distribuiti più equamente (ad esempio, 50% uomini, 50% donne).

#### 3.2 ****In-processing durante l'Addestramento del Modello****

Durante l'addestramento, Fairway monitora il modello e lo corregge in tempo reale se si rilevano decisioni che perpetuano il bias.

**Esempio**: Supponiamo che, nonostante il pre-processing, il modello stia ancora dando preferenza agli uomini per un prestito. Durante l'addestramento, Fairway potrebbe intervenire per modificare il processo di apprendimento del modello, facendogli capire che essere uomo o donna non dovrebbe influenzare la decisione di approvare un prestito.

### 4. ****Equilibrio tra Accuratezza e Equità****

Un problema comune nei modelli di machine learning è il **compromesso tra accuratezza e equità**. Quando si cerca di rendere un modello più equo, spesso si rischia di **ridurre la sua accuratezza** (ovvero la capacità del modello di fare previsioni corrette).

Gli autori, ispirati dal lavoro di Berk et al., usano tecniche di **ottimizzazione multi-obiettivo**. Ciò significa che il loro obiettivo è bilanciare due cose:

* **Ridurre il bias** del modello (aumentando l'equità);
* **Mantenere alte le prestazioni** del modello (senza compromettere troppo l'accuratezza).

**Esempio**: Nel contesto di approvazione dei prestiti, potremmo voler rendere il modello equo in modo che non discrimini sulla base del genere. Tuttavia, vogliamo anche che il modello continui a fare previsioni precise su chi è idoneo a ricevere un prestito, evitando di approvare prestiti a persone con alto rischio di default. Fairway cerca di ottimizzare entrambi gli aspetti, riducendo il bias ma mantenendo alta la qualità delle previsioni.

### 5. ****Risultati: Fairway Funziona Senza Compromettere le Prestazioni****

Gli autori hanno testato Fairway su vari dataset e hanno scoperto che è possibile **mitigare il bias** senza **compromettere in modo significativo le prestazioni del modello**. In altre parole, Fairway riesce a rendere il modello più equo (riducendo la discriminazione) pur mantenendo buone prestazioni predittive.

**Esempio**: Supponiamo che un modello inizialmente avesse un'accuratezza del 90% nell'approvazione dei prestiti, ma fosse altamente sbilanciato a favore degli uomini. Dopo l'uso di Fairway, l'accuratezza potrebbe scendere lievemente (ad esempio, all'88%), ma il modello diventerebbe molto più equo, approvando prestiti in modo più bilanciato tra uomini e donne.

### 6. ****Altri Contributi Importanti del Paper****

* **Individuazione del Bias nei Dati**: Gli autori non si concentrano solo sull'identificazione del bias nel modello, ma anche nei **dati di addestramento**. Capire da dove proviene il bias è fondamentale per correggerlo in modo efficace.
* **Combinazione di Tecniche**: Fairway è unico perché combina approcci di pre-processing e in-processing, rendendolo più potente rispetto a tecniche che si concentrano su una sola fase.
* **Applicabilità Multi-Dominio**: Gli autori discutono come l'equità possa essere definita diversamente a seconda del contesto. Ad esempio, l'equità in ambito giudiziario potrebbe avere implicazioni diverse rispetto a quella in ambito finanziario. Quindi, il concetto di equità è flessibile e deve essere adattato ai vari contesti.

### 7. ****Conclusione****

In conclusione, gli autori propongono che il testing per il bias e la sua mitigazione dovrebbero diventare una **pratica standard** nello sviluppo di software di machine learning. Il metodo **Fairway** permette di individuare e mitigare il bias, migliorando l'equità del modello senza ridurre significativamente le sue prestazioni predittive. Gli autori mettono anche a disposizione il loro strumento su GitHub, in modo che altri ricercatori possano replicare o migliorare i loro risultati.

### Riassunto Finale con Esempio

In breve, gli autori hanno sviluppato un metodo per **identificare e correggere il bias** nei modelli di machine learning, partendo dall'analisi dei dati di addestramento e monitorando l'addestramento del modello stesso. Hanno dimostrato che è possibile rendere un modello più equo senza comprometterne eccessivamente le prestazioni.

**Esempio riassuntivo**:

* Problema: Un modello di approvazione prestiti preferisce approvare prestiti agli uomini piuttosto che alle donne.
* Soluzione di Fairway: Prima corregge i dati sbilanciati (pre-processing), poi interviene durante l'addestramento del modello (in-processing), bilanciando accuratezza e equità.
* Risultato: Il modello diventa più equo nell'approvazione dei prestiti per uomini e donne, mantenendo quasi la stessa accuratezza.

### Spiegazione dettagliata del Background e delle Soluzioni Proposte dagli Autori

In questa sezione del paper, gli autori discutono i concetti fondamentali legati all'**equità nel software** e forniscono una panoramica sui lavori precedenti che trattano il problema del **bias nei modelli di machine learning** (ML). Vediamo i punti principali.

### 2.1 ****Equità del Software****

Gli autori citano vari esempi di **bias nei sistemi software** basati su machine learning:

* **Google Sentiment Analyzer**: Dava un punteggio negativo a frasi come "I am a Jew" o "I am homosexual".
* **Software di riconoscimento facciale**: Classificava erroneamente persone con la pelle scura come animali.
* **Sistemi di traduzione automatica**: Le frasi di genere neutro, come "Lei è un ingegnere, Lui è un infermiere", quando tradotte in turco e poi ritradotte in inglese, diventavano "Lui è un ingegnere, Lei è un'infermiera", mostrando un bias di genere.
* **YouTube Captions**: Gli algoritmi di sottotitolazione automatica erano più precisi con le voci maschili rispetto a quelle femminili.

Questi esempi mostrano che i sistemi basati su machine learning possono **riprodurre o amplificare i bias sociali** presenti nei dati con cui vengono addestrati.

#### ****Problema Principale****:

* Un software è considerato **equo** se non dà vantaggi ingiusti a un determinato gruppo (es. in base a razza, sesso, età). Gli autori sostengono che l'equità dovrebbe essere considerata un aspetto fondamentale della **qualità del software**, al pari di altre caratteristiche come l'efficienza o la sicurezza.

**Esempio**: Un sistema di raccomandazione di prestiti bancari che favorisce sistematicamente gli uomini rispetto alle donne, anche a parità di qualifiche, è un esempio di software iniquo.

### 2.2 ****Lavori Precedenti****

Gli autori forniscono una panoramica dei lavori esistenti sul tema dell'equità nel machine learning e nell'ingegneria del software:

* **IEEE**, l'**Unione Europea** e **Microsoft** hanno pubblicato principi etici per l'intelligenza artificiale che richiedono che i sistemi siano **equi**.
* IBM ha lanciato un toolkit chiamato **AI Fairness 360** per rilevare e mitigare il bias nei modelli di machine learning.
* **Microsoft** ha creato il gruppo di ricerca FATE (Fairness, Accountability, Transparency, and Ethics in AI).
* **Facebook** ha sviluppato un tool chiamato **Fairness Flow** per verificare se un algoritmo di machine learning è influenzato dal bias.

Questi lavori mostrano come l'**equità nel software** stia diventando un'area di ricerca di crescente interesse, non solo per la comunità accademica, ma anche per le grandi aziende tecnologiche.

### 2.3 ****Individuazione del Bias Etico****

Diversi strumenti e metodologie sono stati sviluppati per **trovare bias etici** nei modelli di machine learning:

* **THEMIS**: Un tool per testare il comportamento discriminatorio dei modelli attraverso test randomizzati sugli attributi. Seleziona valori casuali per verificare se il modello discrimina tra individui.
* **AEQUITAS**: Un altro tool che genera automaticamente input discriminatori per identificare violazioni di equità.
* **IBM Research AI**: Propone un metodo per i modelli "black-box" (modelli non interpretabili) che combina tecniche di **dynamic symbolic execution** (esecuzione simbolica dinamica) e spiegazioni locali per generare casi di test.

Gli autori sottolineano che molti di questi lavori si concentrano solo sul rilevamento del bias nei **modelli già addestrati**. Invece, loro vogliono anche capire come il **bias nei dati di addestramento** influisce sul comportamento del modello, e come affrontare il problema sia nel **pre-processing** dei dati che nell'**in-processing** (durante l'addestramento del modello).

#### ****Esempio****:

Se un modello di machine learning discrimina sistematicamente nei confronti delle persone di colore durante la valutazione del rischio di recidiva, è importante capire se il bias deriva dai **dati di addestramento** o dal modo in cui il modello è stato addestrato.

### 2.4 ****Rimozione del Bias Etico****

Gli autori suddividono i lavori esistenti in tre categorie di tecniche per rimuovere il bias:

1. **Pre-processing**: Intervenire sui dati prima che vengano usati per addestrare il modello, modificandoli per ridurre il bias.
   * **Esempio**: Il metodo **Reweighing**, che assegna pesi diversi agli esempi di addestramento in base alla combinazione gruppo/etichetta per rendere i dati più equi. Se un gruppo è sottorappresentato, viene assegnato un peso maggiore ai suoi esempi.
   * **Ottimizzazione pre-processing**: Calmon et al. propongono un metodo che modifica etichette e caratteristiche in modo probabilistico per garantire equità.
2. **In-processing**: Intervenire durante l'addestramento del modello.
   * **Esempio**: Il metodo **Adversarial Debiasing** di Zhang et al. crea un modello in cui un avversario cerca di determinare un attributo protetto (come il genere) dalle previsioni del modello. Se l'avversario non riesce a individuare l'attributo, il modello viene considerato più equo.
   * **Prejudice Remover**: Kamishima et al. aggiungono un termine di regolarizzazione per ridurre la discriminazione durante l'addestramento del modello.
3. **Post-processing**: Correggere le decisioni del modello dopo l'addestramento.
   * **Esempio**: **Reject Option Classification**: In questo metodo, vengono dati esiti favorevoli ai gruppi svantaggiati e sfavorevoli ai gruppi privilegiati, entro un margine di incertezza attorno al confine decisionale del modello.
   * **Equalized Odds Post-Processing**: Si concentra su metriche specifiche come la **Equal Opportunity Difference** per bilanciare le opportunità per diversi gruppi.

Gli autori notano che Fairway combina le tecniche di **pre-processing** e **in-processing**, evitando così la necessità di interventi post-processing, che possono essere complessi e richiedere una conoscenza approfondita del dominio.

### Conclusione del Background

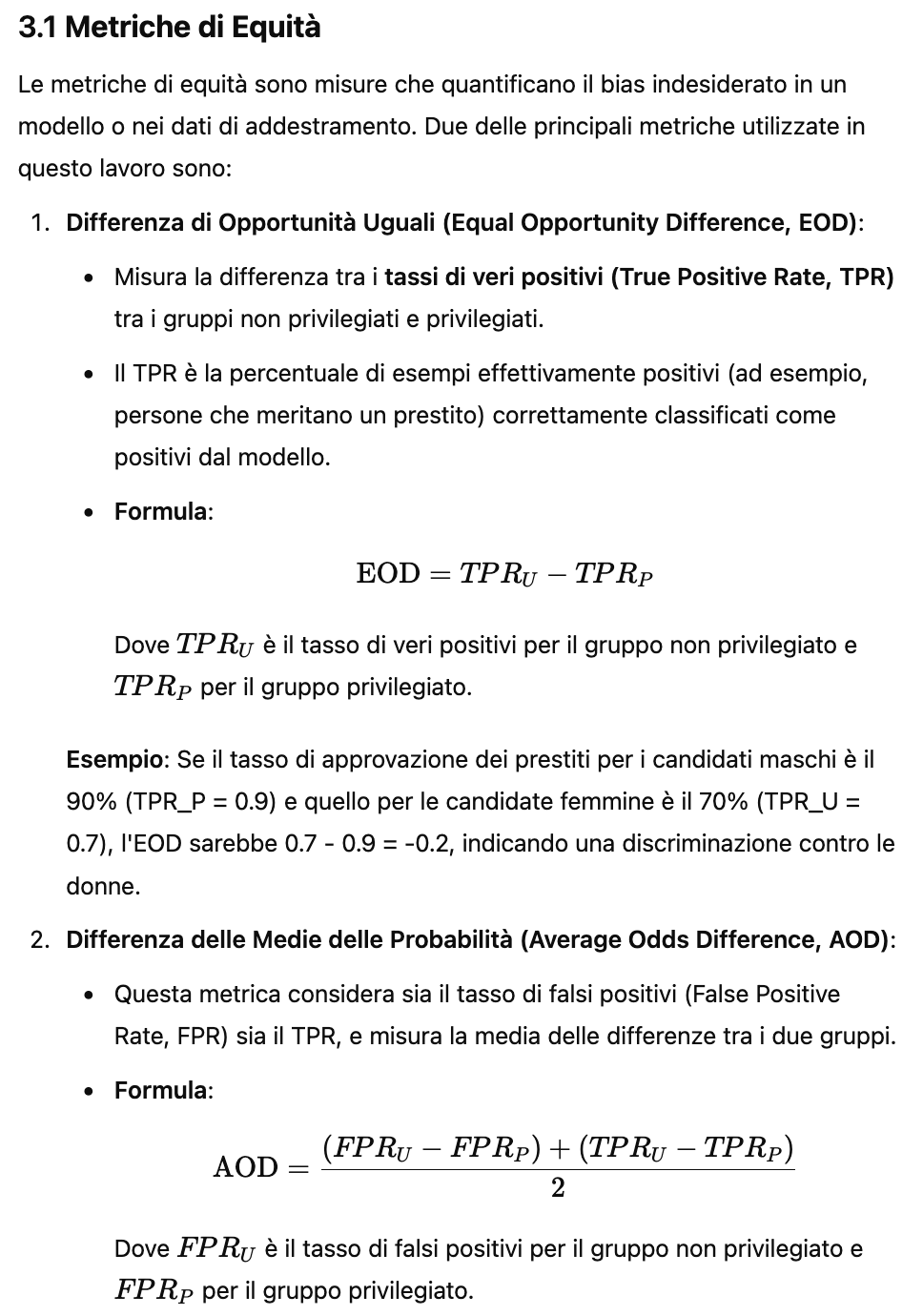
Gli autori propongono che, per costruire software di machine learning equi, è necessario non solo rilevare il bias nei modelli, ma anche intervenire **sia nei dati di addestramento (pre-processing)** che durante l'**addestramento del modello (in-processing)**. Questo approccio integrato permette di mitigare il bias senza compromettere le prestazioni del modello.

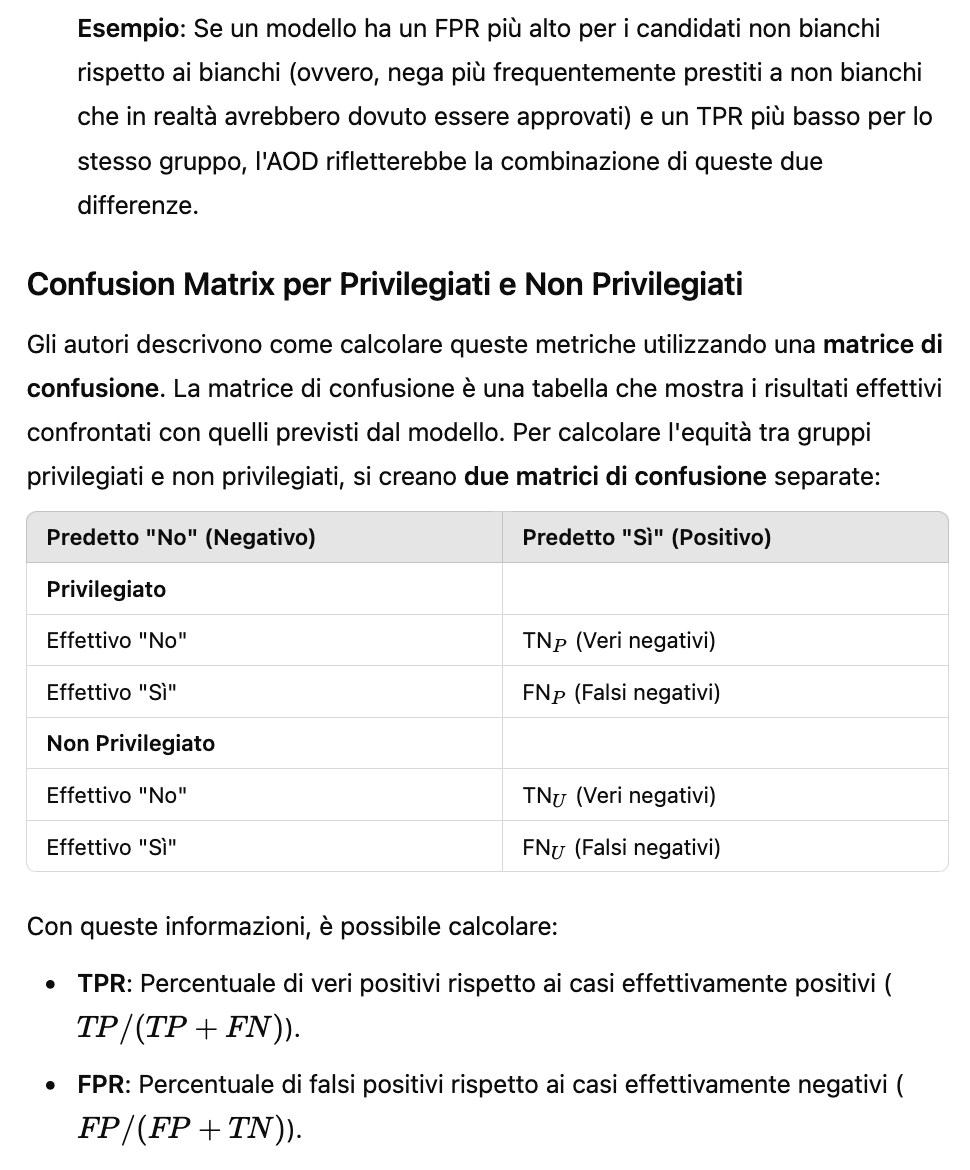
### 3. ****Terminologia dell'Equità****

In questa sezione, gli autori introducono la terminologia chiave utilizzata nel campo dell'**equità nel machine learning**, con particolare attenzione ai modelli di **classificazione binaria** (quelli che prendono decisioni "sì" o "no", come l'approvazione di un prestito) e ai dati in formato tabulare (dati organizzati in righe e colonne). Vediamo i concetti principali descritti in dettaglio.

### Concetti chiave:

1. **Etichetta favorevole**:
   * Un'etichetta favorevole è un risultato positivo per chi riceve una decisione.
   * **Esempio**: Se un modello decide chi riceve un prestito, "approvato" è un'etichetta favorevole perché offre un vantaggio alla persona.
2. **Attributo protetto**:
   * Un attributo protetto è una caratteristica che divide una popolazione in due gruppi: **privilegiato** e **non privilegiato**. Gli attributi protetti sono specifici per l'applicazione.
   * **Esempi comuni di attributi protetti**: "sesso" (maschio/femmina), "razza" (bianco/non-bianco).
3. **Equità di gruppo**:
   * L'obiettivo dell'equità di gruppo è che i gruppi privilegiati e non privilegiati vengano trattati allo stesso modo in base agli attributi protetti. Ad esempio, un modello equo dovrebbe trattare allo stesso modo persone di diverse etnie o generi.
4. **Equità individuale**:
   * L'obiettivo dell'equità individuale è che persone simili ricevano risultati simili, indipendentemente dal gruppo a cui appartengono. Se due persone hanno profili simili, dovrebbero avere le stesse opportunità, indipendentemente dal loro sesso o razza.

****

****

### 3.2 ****Limitazioni delle Metriche di Equità****

Gli autori evidenziano tre limiti delle metriche di equità comunemente usate come AOD ed EOD:

1. **Distribuzione di Classe**:
   * Queste metriche non tengono conto della distribuzione di classi (negativa o positiva) tra gruppi privilegiati e non privilegiati.
   * **Esempio**: Nel **Adult Dataset**, utilizzato per predire il reddito, la distribuzione di classi (persone con basso reddito e persone con alto reddito) varia tra i sessi e le razze. La figura 1 del paper mostra come i gruppi privilegiati (maschi e bianchi) abbiano più casi di reddito alto rispetto ai gruppi non privilegiati (femmine e non bianchi).
2. **Campionamento**:
   * Le metriche non considerano eventuali problemi di **campionamento scorretto** nei dati, che possono creare sbilanciamenti e quindi influire sulla misurazione del bias.
3. **Costo della Misclassificazione**:
   * Le metriche non tengono conto del **costo associato agli errori di classificazione**. Alcuni errori sono più costosi di altri.
   * **Esempio**: In un sistema di valutazione del credito, assegnare un punteggio di credito negativo a un buon cliente è meno costoso che assegnare un punteggio di credito positivo a un cattivo cliente.

### Superamento delle Limitazioni

Gli autori cercano di superare alcune di queste limitazioni:

* **Random Sampling**: Per affrontare i problemi di campionamento scorretto, utilizzano campionamenti casuali dei dati in più ripetizioni.
* **Situation Testing**: Utilizzano un metodo chiamato **situation testing**, che verifica se la previsione del modello cambia quando viene modificato un attributo protetto (ad esempio, cambiando il sesso da maschio a femmina per lo stesso esempio).

Tuttavia, gli autori riconoscono che non affrontano il problema del costo della misclassificazione, poiché richiede una conoscenza specifica del dominio di applicazione.

### Conclusione

In questa sezione, gli autori hanno introdotto le metriche di equità utilizzate per misurare il bias nei modelli di machine learning e hanno discusso le loro limitazioni. Hanno anche spiegato come il loro approccio cerca di migliorare la misurazione dell'equità superando alcune di queste limitazioni, concentrandosi sul confronto tra gruppi privilegiati e non privilegiati.

### 4. ****Descrizione dei Dataset****

Gli autori utilizzano cinque dataset dal repository di machine learning di UC Irvine, che sono spesso utilizzati in studi sull'equità. Ogni dataset ha attributi protetti, ovvero caratteristiche che dividono la popolazione in gruppi privilegiati e non privilegiati. Qui vengono descritti brevemente i dataset utilizzati:

1. **Adult Census Income**:
   * **Descrizione**: Contiene dati su 48.842 persone. L'obiettivo è predire se una persona guadagna più di 50.000 dollari all'anno.
   * **Attributi protetti**: Sesso e razza.
2. **COMPAS**:
   * **Descrizione**: Include dati relativi alla storia criminale e demografica di 7.214 imputati della Contea di Broward, con predizioni del punteggio di rischio COMPAS.
   * **Attributi protetti**: Sesso e razza.
3. **German Credit Data**:
   * **Descrizione**: 1.000 record di persone, con l'obiettivo di predire se una persona ha un "buon" o "cattivo" credito.
   * **Attributo protetto**: Sesso.
4. **Default Credit**:
   * **Descrizione**: 30.000 record di pagamenti inadempienti di persone a Taiwan, con predizioni sul fatto che qualcuno sarà inadempiente o meno.
   * **Attributo protetto**: Sesso.
5. **Heart Health**:
   * **Descrizione**: 297 adulti, con l'obiettivo di predire se una persona ha una condizione cardiaca.
   * **Attributo protetto**: Nessuno specificato, ma altri studi potrebbero considerare età o sesso come attributi protetti.

I dataset vengono utilizzati per valutare l'algoritmo **Fairway** con modelli di classificazione binaria, come la regressione logistica, ma il metodo è applicabile anche ad altri modelli di classificazione.

### 5. ****Metodo Fairway****

**Fairway** è un algoritmo che combina **pre-processing** e **in-processing** per ridurre il bias nei modelli di machine learning. Gli autori spiegano in dettaglio i vari passi coinvolti nel loro metodo, analizzando anche approcci alternativi testati prima di arrivare a Fairway.

#### 5.1 Perché non rimuovere gli attributi protetti?

Uno degli approcci più semplici per mitigare il bias consiste nel **rimuovere gli attributi protetti** (come sesso o razza) dai dati di addestramento e test, in modo che il modello non possa discriminare in base a tali informazioni. Gli autori provano questa strategia sui loro dataset, ma scoprono che non cambia significativamente le metriche di bias.

**Motivo**:

* Anche se l'attributo protetto viene rimosso, **correlazioni tra attributi** possono far sì che il modello continui a discriminare.
* **Esempio**: Nel caso di Amazon, un modello per il servizio di consegna a domicilio era discriminatorio nei confronti di quartieri neri, anche se l'attributo "razza" non era stato utilizzato. Questo perché l'attributo "codice postale" era altamente correlato con la razza, e il modello ha imparato a discriminare indirettamente.

Gli autori esaminano se la correlazione tra attributi sia il motivo per cui il bias persiste nei loro dataset, ma concludono che **il bias proviene principalmente dalle etichette di classe**, ossia da dati storicamente sbagliati o influenzati dal bias umano.

#### 5.2 Rimozione dei Dati Ambigui (Bias)

Per risolvere il problema del bias, Fairway si concentra sulla **rimozione dei punti di dati ambigui** (cioè, quelli che causano bias) dai dati di addestramento. Questo approccio prevede la costruzione di **due modelli separati** per i gruppi privilegiati e non privilegiati (basati sull'attributo protetto), e quindi la verifica se le predizioni di questi modelli siano concordi per ciascun punto dati.

* Se entrambi i modelli concordano su un punto dati, questo viene considerato **non ambiguo** (non biasato) e viene mantenuto.
* Se i modelli non concordano, il punto dati viene rimosso perché potrebbe essere fonte di bias.

**Esempio**: Se un modello predice che una persona nera ha più probabilità di recidivare rispetto a un modello che usa dati su persone bianche con caratteristiche simili, quel dato potrebbe essere considerato ambiguo e rimosso.

Questa pulizia dei dati permette di **ridurre il bias** mantenendo la maggior parte dei dati utili (fino al 15% dei dati viene rimosso nel peggiore dei casi).

#### 5.3 Cosa succede se ci sono due attributi protetti?

Nel caso di dataset con **due attributi protetti** (come il sesso e la razza), gli autori dividono i dati in **quattro gruppi** (due gruppi privilegiati e due non privilegiati) e applicano lo stesso processo di identificazione e rimozione dei dati ambigui.

**Limite**: Se ci sono più di due attributi protetti, i dati diventano troppo frammentati e difficili da gestire, quindi non esplorano ulteriormente questo aspetto.

#### 5.4 Ottimizzazione del Modello

Dopo la rimozione dei dati ambigui, Fairway applica un algoritmo di **multi-obiettivo** per ottimizzare il modello, bilanciando le prestazioni (come il richiamo e il tasso di falsi allarmi) con le metriche di equità (come **AOD** e **EOD**).

**FAIR\_FLASH**: Gli autori usano una versione modificata di **FLASH**, un ottimizzatore basato su modelli sequenziali (SMBO). Invece di valutare tutte le possibili combinazioni di parametri (che sarebbe molto lento), FLASH utilizza **alberi di decisione** per fare previsioni rapide su come potrebbero comportarsi diverse configurazioni di parametri. Questo permette di trovare un compromesso ottimale tra prestazioni ed equità in modo più efficiente.

**Processo**:

1. Dividono il dataset in tre parti: addestramento (70%), validazione (15%) e test (15%).
2. Durante il pre-processing, rimuovono i dati ambigui dal set di addestramento.
3. Addestrano un modello di **regressione logistica** sul set di addestramento.
4. Utilizzano FAIR\_FLASH per trovare i migliori parametri del modello, ottimizzando richiamo, falsi allarmi, AOD ed EOD.
5. Infine, applicano il modello ottimizzato al set di test per la valutazione finale.

### Conclusione

Fairway combina la rimozione del bias dai dati di addestramento con l'ottimizzazione del modello per bilanciare equità e prestazioni. Questo approccio si è dimostrato efficace, perdendo solo una piccola parte dei dati (fino al 15%) e mantenendo elevate le prestazioni del modello. Fairway è particolarmente utile per dataset con uno o due attributi protetti, ma potrebbe richiedere ulteriori studi per gestire casi con più attributi protetti o dataset molto complessi.

### 6. ****Risultati****

I risultati dell'articolo sono organizzati attorno a sei domande di ricerca, con lo scopo di valutare l'efficacia del metodo **Fairway** nel rilevare e mitigare il bias nei modelli di machine learning. Gli esperimenti sono stati ripetuti dieci volte con rimescolamenti dei dati, e i risultati riportano i valori mediani.

### ****RQ1: Quali sono i problemi con i modelli standard?****

Gli autori dimostrano che i metodi di apprendimento standard, come la **regressione logistica** con parametri predefiniti, presentano problemi di equità. Hanno testato questo approccio sui cinque dataset, con metriche come richiamo (recall), allarmi falsi (false alarm), AOD (Average Odds Difference), e EOD (Equal Opportunity Difference). Le metriche di equità (AOD ed EOD) mostrano che i modelli standard non sono equi, favorendo i gruppi privilegiati (es. uomini o bianchi) rispetto ai gruppi non privilegiati (es. donne o non bianchi).

**Risultato**: I modelli standard non sono equi e privilegiano determinati gruppi, fornendo un punto di riferimento di partenza per l'esperimento.

### ****RQ2: Quanto migliora il pre-processing i risultati?****

In questa domanda, gli autori testano l'efficacia della rimozione dei dati ambigui (pre-processing). Dopo aver rimosso i dati potenzialmente biasati, il modello è stato ritrainato.

* **Risultati**: I risultati mostrano un leggero deterioramento nel richiamo, ma miglioramenti significativi nelle metriche di equità (AOD, EOD). Questo indica che il pre-processing dei dati migliora notevolmente l'equità del modello, riducendo il bias.

### ****RQ3: Quanto migliora l'ottimizzazione dei risultati?****

Qui gli autori verificano l'efficacia dell'ottimizzazione del modello usando **FAIR\_FLASH**, senza il pre-processing. L'ottimizzazione viene fatta per migliorare il richiamo e ridurre AOD ed EOD.

* **Risultati**: L'ottimizzazione migliora il richiamo e gli allarmi falsi, ma non è così efficace nel ridurre AOD ed EOD rispetto al solo pre-processing. Questo suggerisce che l'ottimizzazione da sola è utile, ma il pre-processing aggiunge un valore significativo.

### ****RQ4: Quanto migliora Fairway i risultati?****

La quarta domanda esplora l'effetto combinato di pre-processing e ottimizzazione (cioè l'intero metodo **Fairway**).

* **Risultati**: Fairway migliora significativamente le metriche di equità in quasi tutti i casi, mantenendo il richiamo simile all'ottimizzazione. Ad esempio:
  + Nel dataset **Adult**, per l'attributo protetto "razza", Fairway ottiene un richiamo simile all'ottimizzazione, ma con un notevole miglioramento nelle metriche di equità.
  + Nel dataset **Default Credit**, per l'attributo "sesso", Fairway fornisce i migliori risultati su tutte le metriche.

Fairway riesce a migliorare l'equità senza danneggiare significativamente le prestazioni del modello, dimostrando l'importanza di combinare pre-processing e ottimizzazione.

### ****RQ5: Come si confronta Fairway con gli algoritmi precedenti di mitigazione del bias?****

Gli autori confrontano Fairway con cinque algoritmi di mitigazione del bias precedenti, presi dalla libreria IBM AIF360, e applicano questi algoritmi a tre dataset: Adult, COMPAS, e German.

* **Risultati**: Fairway riduce il bias in modo simile o migliore rispetto agli algoritmi precedenti in termini di AOD ed EOD, e danneggia il richiamo meno frequentemente rispetto agli altri approcci. Quando Fairway perde nel richiamo, la perdita è minima (10-12%).

Fairway si distingue perché combina il rilevamento del bias e la sua mitigazione, mentre gli algoritmi precedenti si concentrano principalmente sulla mitigazione.

### ****RQ6: Può Fairway essere usato come strumento combinato per il rilevamento e la mitigazione del bias?****

L'ultima domanda esplora l'uso di Fairway come uno strumento combinato per rilevare e mitigare il bias. Gli autori applicano **situation testing**, verificando se il modello cambia la sua predizione quando un attributo protetto (es. sesso) viene modificato.

* **Risultati**: Fairway riduce significativamente la percentuale di dati che falliscono il situation testing (ossia i dati che mostrano bias), mostrando che può essere usato non solo per mitigare il bias, ma anche per **rilevarlo** nei dati di addestramento. Fairway può quindi essere utilizzato per verificare e correggere il bias nei modelli senza rivelare i dettagli interni del modello, rendendolo utile anche in contesti industriali.

### ****Sintesi dei Risultati****

* **RQ1**: I modelli standard sono spesso ingiusti.
* **RQ2**: Il pre-processing migliora l'equità.
* **RQ3**: L'ottimizzazione migliora le prestazioni, ma meno l'equità.
* **RQ4**: Fairway (pre-processing + ottimizzazione) offre il miglior compromesso tra prestazioni ed equità.
* **RQ5**: Fairway compete bene con gli algoritmi precedenti, danneggiando meno le prestazioni.
* **RQ6**: Fairway può essere usato sia per rilevare che per mitigare il bias.

Fairway si dimostra un approccio efficace per bilanciare l'equità senza compromettere significativamente le prestazioni, e rappresenta uno strumento flessibile per rilevare e mitigare il bias nei modelli di machine learning.

### 7. ****Minacce alla Validità****

Gli autori evidenziano quattro principali minacce alla validità dei loro risultati, fornendo una riflessione critica sui limiti del loro lavoro e proponendo futuri miglioramenti.

1. **Bias di Campionamento**:
   * **Descrizione**: I cinque dataset utilizzati provengono dal repository di machine learning dell'Università della California a Irvine (UCI), che è comunemente utilizzato nelle ricerche sull'equità. Tuttavia, c'è una preoccupazione riguardo alla generalizzabilità dei risultati ottenuti da questi dataset a contesti più ampi.
   * **Minaccia**: Questi dataset potrebbero non riflettere una gamma sufficientemente ampia di problemi di equità che si potrebbero incontrare in altri dataset.
   * **Mitigazione futura**: Gli autori propongono di esplorare un numero maggiore di dataset in ricerche future per verificare se i risultati si estendono ad altre tipologie di dati.
2. **Bias di Valutazione**:
   * **Descrizione**: Gli autori hanno utilizzato due metriche di equità: **EOD** (Equal Opportunity Difference) e **AOD** (Average Odds Difference). Tuttavia, queste metriche non tengono conto della distribuzione delle classi, concentrandosi solo su TPR (True Positive Rate) e FPR (False Positive Rate).
   * **Minaccia**: Le metriche utilizzate possono non essere sufficienti per valutare adeguatamente l'equità del modello.
   * **Mitigazione futura**: Gli autori citano metriche alternative, come la **Conditional Equality of Opportunity**, che potrebbero superare questa limitazione considerando la distribuzione delle caratteristiche. Propongono di esplorare queste metriche in studi futuri.
3. **Validità del Costrutto**:
   * **Descrizione**: Gli autori hanno addestrato modelli separati per i gruppi privilegiati e non privilegiati basandosi su uno o due attributi protetti per ogni dataset. Questo approccio ha funzionato con i dataset utilizzati, che contenevano al massimo due attributi protetti.
   * **Minaccia**: La metodologia potrebbe non essere scalabile per dataset più complessi con più attributi protetti.
   * **Mitigazione futura**: In futuro, gli autori propongono di esplorare dataset più grandi e complessi con più attributi protetti per verificare la fattibilità e l'efficacia del loro approccio.
4. **Validità Esterna**:
   * **Descrizione**: Il metodo **Fairway** è stato applicato solo a modelli di **classificazione**, che sono molto comuni nei software di intelligenza artificiale (AI).
   * **Minaccia**: Fairway non è stato testato su altri tipi di problemi di machine learning, come la **regressione** o l'elaborazione di testi e video.
   * **Mitigazione futura**: Gli autori stanno lavorando per estendere Fairway ai modelli di regressione e ad altri problemi di data mining, come l'elaborazione di testi e video, ampliando così il campo di applicazione del loro metodo.

### 8. ****Conclusione****

In questa sezione conclusiva, gli autori riassumono i principali contributi del loro lavoro e delineano le prospettive future.

* **Bias nei Dati di Addestramento**: Gli autori hanno spiegato come i modelli di machine learning acquisiscano bias da etichette errate o distorte nei dati di addestramento. Questo bias si riflette poi nelle predizioni dei modelli, perpetuando l'**inequità**.
* **Fairway**: Hanno proposto un approccio chiamato **Fairway**, che non solo mitiga il bias etico, ma rileva anche bias nei dati e ottimizza il modello per massimizzare l'equità e le prestazioni. Fairway è un metodo combinato di pre-elaborazione e ottimizzazione che rimuove i dati ambigui (che causano bias) e ottimizza il modello per ridurre il bias.
* **Contributo Pratico**: Fairway è comprensibile e può essere utilizzato come strumento sia per rilevare che per mitigare il bias, rendendo più equi i modelli di machine learning. Gli autori hanno reso il codice di Fairway disponibile pubblicamente per la comunità di ricerca e per i professionisti del software, rendendolo uno strumento accessibile.
* **Differenze con i Lavori Precedenti**: Fairway si differenzia dagli approcci precedenti perché non si limita a mitigare il bias, ma convalida anche le etichette dei dati e rileva i dati distorti prima di addestrare il modello. È un approccio completo che combina rilevamento e mitigazione del bias in un unico strumento.
* **Prospettive Future**: Gli autori auspicano che più ricercatori e industrie lavorino in questo campo e che vengano pubblicati più dataset per ampliare le possibilità di studio. Con l'accesso a più dati, sarebbe opportuno ripetere questo studio e valutarne la generalizzabilità.

In sintesi, **Fairway** è un metodo innovativo e completo per la rilevazione e la mitigazione del bias nei modelli di machine learning. Questo studio segna un passo avanti verso la creazione di modelli di AI più equi e giusti, e pone le basi per future ricerche in questo ambito cruciale.

**RECAP**

Gli autori di questo paper hanno sviluppato un metodo chiamato **Fairway** per rilevare e mitigare il bias etico nei modelli di machine learning, in particolare nei modelli di **classificazione binaria**. Il bias etico si verifica quando un modello di machine learning tratta in modo ingiusto determinati gruppi di persone, come uomini vs. donne, bianchi vs. neri, ecc., a causa di **dati di addestramento distorti** o **etichette improprie**. Qui ti spiego passo per passo cosa hanno fatto gli autori.

### 1. ****Problema che stanno affrontando****

Il bias nei modelli di machine learning è un problema perché può portare a decisioni discriminatorie. Ad esempio, un modello potrebbe assegnare tassi di interesse più bassi agli uomini rispetto alle donne, anche se le donne hanno le stesse caratteristiche di rischio. Gli autori vogliono:

1. Rilevare dove si trova questo bias nei dati di addestramento.
2. Rimuovere il bias senza compromettere troppo le prestazioni predittive del modello.

### 2. ****Dataset Utilizzati****

Hanno utilizzato 5 dataset ben noti che presentano casi di **bias**:

* **Adult Income Dataset**: Predice se una persona guadagna più di 50k dollari all'anno (bias legato a sesso e razza).
* **COMPAS**: Predice se un imputato sarà recidivo (bias legato a sesso e razza).
* **German Credit Dataset**: Predice se una persona avrà un buon credito (bias legato al sesso).
* **Default Credit**: Predice se una persona sarà inadempiente (bias legato al sesso).
* **Heart Health**: Predice se una persona ha una malattia cardiaca.

### 3. ****Cosa ha fatto Fairway?****

Fairway è un **approccio a due fasi** che combina:

1. **Pre-processing (Rimozione del bias nei dati)**: L'obiettivo qui è identificare i punti dati ambigui o biasati nei dati di addestramento. Ad esempio, se un dato dice che un uomo con determinate caratteristiche ottiene un prestito, ma una donna con le stesse caratteristiche non lo ottiene, Fairway lo considera un **dato ambiguo**.
   * **Come fanno?** Dividono i dati in due gruppi: uno privilegiato (es. uomini, bianchi) e uno non privilegiato (es. donne, non bianchi). Poi addestrano **due modelli separati** su ciascun gruppo e controllano se fanno previsioni diverse sugli stessi punti dati. Se i due modelli danno risultati diversi per lo stesso dato, quel punto viene rimosso perché potrebbe essere fonte di bias.
2. **Ottimizzazione del modello (FAIR\_FLASH)**: Dopo aver rimosso i dati ambigui, ottimizzano il modello di machine learning per bilanciare due obiettivi:
   * Migliorare l'equità (ridurre il bias).
   * Mantenere buone prestazioni predittive. Questo viene fatto utilizzando un algoritmo chiamato **FAIR\_FLASH**, che cerca i migliori parametri del modello per ridurre il bias e mantenere alta la capacità predittiva.

### 4. ****Esperimenti e Risultati****

Gli autori hanno fatto sei esperimenti per verificare quanto fosse efficace Fairway. Ecco cosa hanno scoperto:

1. **Problemi con i modelli standard**: I modelli di machine learning tradizionali, senza mitigazione del bias, mostravano **bias significativi** a favore di gruppi privilegiati.
2. **Effetto del Pre-processing (Rimozione del bias)**: Dopo aver rimosso i dati ambigui, hanno osservato un **miglioramento significativo nell'equità**. Il bias è diminuito senza danneggiare troppo le prestazioni del modello.
3. **Effetto dell'Ottimizzazione**: Quando hanno ottimizzato solo il modello (senza rimuovere i dati ambigui), hanno visto miglioramenti nelle prestazioni del modello, ma il bias è rimasto. Quindi, solo l'ottimizzazione non risolve il problema del bias.
4. **Effetto della combinazione (Fairway)**: Combinando la rimozione dei dati ambigui e l'ottimizzazione, Fairway ha ottenuto i **migliori risultati**: alta equità e buone prestazioni del modello.
5. **Confronto con altri algoritmi di mitigazione**: Fairway è stato confrontato con altri algoritmi noti per la mitigazione del bias. Si è dimostrato migliore o simile nella riduzione del bias, con meno impatto negativo sulle prestazioni predittive.
6. **Fairway come strumento combinato**: Fairway non solo mitiga il bias, ma è anche utile per **rilevare** dove si trova il bias nei dati di addestramento. Inoltre, può essere utilizzato con modelli di machine learning di tipo "black-box" (cioè modelli di cui non si conosce la logica interna).

### 5. ****Conclusioni****

* **Fairway** non è solo uno strumento di mitigazione del bias, ma un approccio completo per rilevare e rimuovere il bias dai dati di addestramento prima ancora di addestrare un modello.
* È efficace nel bilanciare l'equità senza sacrificare troppo le prestazioni predittive.
* Gli autori hanno reso il codice di Fairway disponibile al pubblico, in modo che altri possano usarlo o migliorarlo.

### Esempio di utilizzo di Fairway:

Immagina di avere un dataset che predice se una persona ottiene un prestito. Se ci sono più uomini che ottengono prestiti rispetto alle donne con le stesse caratteristiche economiche, Fairway può rilevare questi **dati distorti** e rimuoverli prima di addestrare il modello, garantendo che il modello finale non sia discriminatorio nei confronti delle donne.

Measuring and Mitigating Unintended Bias in Text Classification

2019

In questo articolo, gli autori presentano un approccio per **misurare e mitigare il bias non intenzionale** nei modelli di machine learning, focalizzandosi sui **classificatori di testo**. Questo metodo è applicato a un corpus pubblico di commenti annotati per la tossicità provenienti dalle pagine di discussione di Wikipedia. Ecco una spiegazione dettagliata di cosa hanno fatto gli autori e come hanno affrontato il problema del bias nei loro modelli:

### 1. ****Problema del Bias nei Modelli di Machine Learning****

Il bias nei modelli di machine learning può essere il risultato di **squilibri nei dati di addestramento**. Se il modello è addestrato su dati che riflettono pregiudizi umani, c'è il rischio che il modello stesso perpetui questi pregiudizi, generando prestazioni ingiuste. In questo contesto, i modelli potrebbero, ad esempio, **classificare erroneamente** alcune frasi come tossiche semplicemente perché contengono termini associati a gruppi demografici, come “gay” o “nero”, anche se il contenuto della frase non è offensivo.

### 2. ****Esempio di Bias nel Classificatore di Testo****

Nel loro studio, i ricercatori si sono concentrati su un **classificatore di testo** addestrato per rilevare commenti tossici nelle pagine di discussione di Wikipedia. Hanno notato che frasi come "I am a gay man" (che sono chiaramente non tossiche) venivano spesso classificate come tossiche a causa dell'uso frequente di termini identitari come "gay" in commenti tossici. Questo fenomeno viene definito **false positive bias** (falsi positivi dovuti al bias).

### 3. ****Origine del Bias****

Il bias rilevato derivava dalla **sproporzionata rappresentazione** di termini identitari nei dati di addestramento. Ad esempio, i termini come "gay" erano spesso associati a commenti tossici, portando il modello a sovra-generalizzare e ad associare automaticamente questi termini a tossicità, anche quando non era il caso.

### 4. ****Definizione di Bias Non Intenzionale****

Gli autori propongono una **definizione di bias non intenzionale** nei modelli di classificazione, che dipende da due fattori:

* Il **set di test** utilizzato per valutare il modello.
* Un **sottoinsieme di caratteristiche di input** (in questo caso, termini demografici specifici come “gay” o “nero”).

L'idea è che se il modello mostra una classificazione non bilanciata basata su queste caratteristiche demografiche (come associare certi termini identitari a tossicità), ciò è considerato un **bias non intenzionale**.

### 5. ****Metodo per Mitigare il Bias****

Per mitigare il bias, gli autori propongono un approccio **non supervisionato** basato sull'equilibrio del dataset di addestramento. L'approccio prevede di **aggiungere dati strategicamente** per bilanciare la rappresentazione dei termini identitari nel dataset di addestramento, evitando così che il modello impari associazioni distorte.

### 6. ****Esempio del Metodo di Mitigazione****

Ad esempio, se il termine "gay" compare principalmente in commenti tossici nei dati di addestramento, il modello potrebbe imparare ad associare il termine "gay" a tossicità. Per bilanciare questa rappresentazione, gli autori suggeriscono di **aggiungere dati** che contengono il termine "gay" in un contesto non tossico, in modo che il modello impari a distinguere tra contesti tossici e non tossici che includono termini identitari.

### 7. ****Metriche di Valutazione****

Gli autori introducono metriche per valutare l'**entità del bias** nei modelli, misurando quanto il modello sia distorto nelle sue previsioni per determinati termini identitari. Hanno anche valutato l'**impatto del loro metodo di mitigazione**, dimostrando che la loro tecnica riduce il bias senza compromettere la **qualità generale del modello**.

### 8. ****Risultati****

Dopo aver applicato il loro metodo di bilanciamento del dataset, i ricercatori hanno osservato che il modello presentava una **riduzione del bias non intenzionale** (come la sovra-associazione del termine "gay" con tossicità) pur mantenendo buone prestazioni complessive nel rilevare commenti tossici.

### Conclusione

In sintesi, gli autori:

* Hanno definito un approccio per **misurare il bias** nei modelli di classificazione del testo, concentrandosi su un sottoinsieme di termini demografici.
* Hanno dimostrato che il bias può derivare da **squilibri nei dati di addestramento**, che portano il modello a fare previsioni distorte.
* Hanno proposto una tecnica per **bilanciare i dati di addestramento** in modo non supervisionato, riducendo il bias senza compromettere la qualità complessiva del modello.
* Hanno applicato e validato il loro metodo su un classificatore di commenti tossici, mostrando come questo approccio possa migliorare l'equità delle predizioni del modello.

**Lavori Correlati (Italian Translation)**  
I ricercatori nel campo dell'equità nell'apprendimento automatico (ML) hanno proposto una vasta gamma di definizioni di "equità" e metriche per la sua valutazione. Molti hanno anche presentato strategie di mitigazione per migliorare l'equità del modello in base a queste metriche. Feldman et al. (2015) fornisce una definizione di equità legata alla parità demografica delle predizioni del modello, e propone una strategia per modificare i dati di addestramento al fine di migliorare l'equità. Hardt, Price e Srebro (2016) presentano una definizione alternativa di equità che richiede la parità nelle prestazioni del modello invece delle predizioni, insieme a una strategia di mitigazione che si applica ai modelli già addestrati. Kleinberg, Mullainathan e Raghavan (2016) e Friedler, Scheidegger e Venkatasubramanian (2016) confrontano diverse metriche di equità. Questi studi si basano sulla disponibilità di dati demografici riguardanti l'oggetto di classificazione per identificare e mitigare i bias. Beutel et al. (2017) presenta una nuova tecnica di mitigazione basata sull'addestramento avversario che richiede solo una piccola quantità di dati demografici etichettati.

Pochi lavori precedenti hanno affrontato l'equità nei compiti di classificazione testuale. Blodgett e O'Connor (2017), Hovy e Spruit (2016) e Tatman (2017) discutono l'impatto dell'uso di modelli di elaborazione del linguaggio naturale ingiusti per compiti nel mondo reale, ma non forniscono strategie di mitigazione. Bolukbasi et al. (2016) dimostra il bias di genere nei word embeddings e propone una tecnica per "de-biasarli", consentendo a questi embeddings più equi di essere utilizzati per qualsiasi compito basato sul testo.

Il nostro lavoro contribuisce a questa crescente letteratura sulla ricerca di equità nell'apprendimento automatico con un approccio innovativo per definire, misurare e mitigare i bias involontari in un compito di classificazione testuale.

Gli autori di questo lavoro si concentrano su come misurare e mitigare il bias non intenzionale nei modelli di machine learning, in particolare per i classificatori di testo utilizzati per identificare commenti tossici. Di seguito ti spiego in dettaglio la metodologia utilizzata dagli autori, con esempi per chiarire ogni passaggio.

### 1. ****Modello e Dati Utilizzati****

Gli autori lavorano con un **classificatore di testo** progettato per rilevare la **tossicità** nei commenti provenienti dalle pagine di discussione di Wikipedia. Il modello è stato addestrato su un dataset di 127.820 commenti etichettati da revisori umani come **tossici** o **non tossici**. Un commento tossico viene definito come "maleducato, irrispettoso o irragionevole, tale da far lasciare la discussione". Il modello utilizzato è una **rete neurale convoluzionale** (CNN) implementata con Keras in TensorFlow.

### 2. ****Definizione di Bias Non Intenzionale e di Equità****

Gli autori chiariscono la distinzione tra:

* **Bias non intenzionale**: Si verifica quando un modello esprime un comportamento di classificazione che non è stato esplicitamente voluto, come discriminare in base al genere o all'identità demografica. Per esempio, se un modello classifica il termine "gay" come tossico, indipendentemente dal contesto.
* **Impatto equo**: Riguarda l'effetto che il modello ha su persone o gruppi nella pratica. Ad esempio, se un sito web usa il modello per rimuovere commenti con punteggi di tossicità elevati, potrebbe rendere difficile la discussione di argomenti che includono il termine "gay", generando un impatto ingiusto.

Quindi, il **bias non intenzionale** riguarda il comportamento del modello (es. assegnare un punteggio tossico a commenti non tossici), mentre l'**impatto equo** è come questo comportamento influisce sugli utenti.

### 3. ****Quantificazione del Bias****

Gli autori hanno rilevato che termini identitari, come "gay", sono usati in modo sproporzionato nei commenti tossici rispetto a quelli non tossici. Ad esempio, la parola "gay" appare nel 3% dei commenti tossici, ma solo nello 0,5% dei commenti in generale. Questa sproporzione nei dati di addestramento fa sì che il modello **sovergeneralizzi**, associando automaticamente il termine "gay" alla tossicità, anche quando non lo è.

### 4. ****Mitigazione del Bias****

Per ridurre questo bias non intenzionale causato dalla sproporzione nei dati, gli autori hanno adottato la seguente strategia:

1. **Aggiunta di dati non tossici**: Hanno aggiunto nuovi esempi di commenti **non tossici** contenenti i termini identitari che erano stati associati in modo sproporzionato alla tossicità. L'obiettivo era riportare l'equilibrio tra commenti tossici e non tossici in linea con la distribuzione complessiva del dataset.
2. **Bilanciamento in base alla lunghezza**: Poiché i commenti tossici tendono a essere più brevi, hanno anche bilanciato i nuovi dati in base alla lunghezza del commento. Questo perché i modelli CNN sono sensibili alla lunghezza del testo, e bilanciare la lunghezza aiuta a evitare che il modello sovra-generalizzi su commenti corti contenenti termini identitari.

### 5. ****Raccolta dei Dati Aggiuntivi****

Per raccogliere questi dati non tossici aggiuntivi, hanno utilizzato i **testi degli articoli di Wikipedia**, assumendo che questi articoli fossero generalmente non tossici. Hanno validato questa ipotesi etichettando manualmente 1000 commenti e trovando che il 99,5% di essi era effettivamente non tossico. Questo metodo è risultato efficace perché:

* Non richiede **etichettatura umana aggiuntiva** (che sarebbe costosa).
* Anche se i testi provenivano da un dominio leggermente diverso (articoli vs. commenti), la differenza non ha compromesso l'efficacia della mitigazione del bias.

### 6. ****Esempio di Mitigazione del Bias****

Supponiamo che il termine "gay" sia associato al 3% dei commenti tossici ma solo allo 0,5% dei commenti non tossici. Per ridurre il bias, gli autori hanno aggiunto **nuovi commenti non tossici** contenenti il termine "gay" fino a riportare il rapporto tra commenti tossici e non tossici contenenti questo termine in linea con quello generale del dataset. Questo riduce la tendenza del modello a sovra-associare il termine "gay" alla tossicità.

### Conclusioni

Gli autori hanno dimostrato che è possibile:

1. **Misurare il bias** nei modelli di classificazione del testo utilizzando termini identitari e analizzando la loro rappresentazione nei dati.
2. **Mitigare il bias** non intenzionale bilanciando il dataset di addestramento, aggiungendo esempi non tossici contenenti quei termini identitari sovra-rappresentati nei commenti tossici.
3. Utilizzare **dati non supervisionati** provenienti da un dominio correlato (come gli articoli di Wikipedia) è una strategia efficace per mitigare il bias senza richiedere costose etichettature manuali.

In sintesi, hanno mostrato come uno squilibrio nei dati di addestramento possa portare a bias non intenzionali e hanno proposto una soluzione pratica ed efficace per ridurre questo problema, mantenendo alta la qualità del modello.

Gli autori hanno introdotto diverse metriche per valutare e mitigare il bias non intenzionale nei modelli di classificazione di testo, in particolare per identificare commenti tossici. Qui ti spiego il metodo e le metriche utilizzate per valutare le prestazioni e il bias nei modelli, con esempi per facilitarne la comprensione.

### 1. ****Test Set Utilizzati****

* **General Test Set**: Questo set è un campione dei commenti originari delle pagine di discussione di Wikipedia ed è usato per valutare le prestazioni complessive del modello. Tuttavia, non è utile per identificare bias non intenzionali, ma serve a garantire che le tecniche di mitigazione del bias non danneggino drasticamente le prestazioni generali.
* **Identity Phrase Templates Test Set**: Questo è un dataset sintetico creato appositamente per testare il bias. Vengono utilizzati **template di frasi** (come "I am <IDENTITY>" o "I hate all <IDENTITY>") nei quali vengono inseriti diversi termini identitari (ad esempio, gay, nero, musulmano). Contiene 77.000 esempi, divisi equamente tra commenti tossici e non tossici.

### 2. ****Metriche di Valutazione****

Le seguenti metriche sono state utilizzate per valutare il bias non intenzionale e le prestazioni dei modelli:

#### A. ****AUC (Area Under the Curve)****

* **Definizione**: L'AUC è una misura comune per valutare le prestazioni di un modello. Essa rappresenta l'area sotto la curva ROC (Receiver Operating Characteristic), che misura quanto bene un modello separa correttamente classi positive e negative.
* **Limiti**: L'AUC da sola non fornisce una valutazione dettagliata del bias. Ad esempio, se si calcola l'AUC per ogni termine identitario (es. "gay", "musulmano", ecc.), si potrebbe ottenere un punteggio alto per ogni sottogruppo, ma questo potrebbe nascondere il bias. È necessario un ulteriore approfondimento per identificare quali sottogruppi sono trattati in modo diverso dal modello.

#### B. ****Error Rate Equality Difference****

* **Definizione**: Questa metrica si basa sul concetto di "Equality of Odds" (Hardt et al., 2016), che si raggiunge quando il modello ha tassi di falsi positivi e falsi negativi simili per diversi gruppi demografici. L'idea è che un modello sia più equo se il tasso di errore per diversi termini identitari è simile.
* **Come si calcola**: Si calcolano il tasso di falsi positivi (FPR) e falsi negativi (FNR) su tutto il test set, e poi si confrontano questi tassi per ogni termine identitario:
  + **False Positive Equality Difference** = ∑ |FPR - FPR\_t|
  + **False Negative Equality Difference** = ∑ |FNR - FNR\_t| Qui, FPR e FNR sono i tassi complessivi, mentre FPR\_t e FNR\_t sono i tassi per un determinato termine identitario (ad esempio, "gay").
* **Interpretazione**: Se c'è molta variazione tra i tassi di errore per i diversi gruppi identitari, significa che il modello ha un alto bias non intenzionale.

#### C. ****Pinned AUC****

* **Problema con l'AUC tradizionale**: Come menzionato, l'AUC calcolata separatamente per ogni termine identitario potrebbe non riflettere il bias non intenzionale. Ad esempio, il modello potrebbe avere un'elevata AUC per ciascun gruppo identitario (come "tall", "short", "average"), ma nascondere il bias nel modo in cui tratta un sottogruppo rispetto agli altri.
* **Pinned AUC**: Questa metrica affronta il problema creando un dataset ausiliario che combina esempi di un sottogruppo specifico (ad esempio, commenti che contengono "gay") con un campione della distribuzione generale dei dati. L'idea è "pinnare" (ancorare) un sottogruppo alla distribuzione complessiva per vedere quanto il modello tratta quel sottogruppo in modo diverso rispetto alla media.

Formalmente:

* + **pDt** = s(Dt) + s(D) (un dataset con metà esempi dal sottogruppo e metà dalla distribuzione generale)
  + **pAUCt** = AUC(pDt) (calcolo dell'AUC per questo dataset bilanciato)

**Esempio**: Se il modello tratta in modo diverso i commenti contenenti il termine "gay", il **pinned AUC** per questo termine sarà diverso dall'AUC complessiva del modello.

#### D. ****Pinned AUC Equality Difference****

* **Definizione**: Questa metrica misura quanto i valori di pinned AUC variano tra i diversi sottogruppi. Idealmente, le pinned AUC per tutti i termini identitari dovrebbero essere simili tra loro e simili all'AUC complessiva.
* **Formula**:
  + Pinned AUC Equality Difference = ∑ |AUC - pAUCt| Dove AUC è l'AUC complessiva e pAUCt è la pinned AUC per un termine identitario specifico. Un valore basso indica che il modello ha una performance simile per tutti i sottogruppi, riducendo così il bias.

### Conclusioni

Gli autori hanno proposto queste metriche per valutare e mitigare il bias non intenzionale nei modelli di classificazione di testo, con particolare attenzione ai termini identitari come "gay", "nero", "musulmano", ecc. L'obiettivo è bilanciare le prestazioni generali del modello con la riduzione del bias nei confronti di specifici gruppi demografici.

Gli esperimenti condotti dagli autori mirano a valutare l'efficacia della tecnica di mitigazione del bias introdotta, confrontando tre modelli:

1. **Modello Baseline**: Addestrato su tutti i 127.820 commenti supervisionati delle pagine di discussione di Wikipedia.
2. **Modello Mitigato**: Addestrato utilizzando la stessa architettura, ma con l'aggiunta di 4.620 nuovi esempi non tossici (assunti tali) per bilanciare la distribuzione dei termini identitari.
3. **Modello di Controllo**: Simile al modello mitigato, ma con l'aggiunta di 4.620 commenti casuali selezionati dagli articoli di Wikipedia, per assicurarsi che i miglioramenti non siano solo dovuti all'aggiunta di dati.

### Valutazione delle Prestazioni dei Modelli

#### 1. ****AUC (Area Under the Curve) Complessiva****

La tabella presentata mostra che il modello mitigato ottiene i migliori risultati nei test con le **template di frasi identitarie**, mantenendo un'elevata performance generale. I valori di AUC sono:

* **Modello Baseline**: 0.960 nel test generale e 0.952 nel test con template identitari.
* **Modello di Controllo**: 0.957 nel test generale e 0.946 nel test con template identitari.
* **Modello Mitigato**: 0.959 nel test generale e 0.960 nel test con template identitari.

Ciò dimostra che la tecnica di mitigazione del bias ha ridotto efficacemente il bias non intenzionale senza compromettere le prestazioni generali del modello.

#### 2. ****Tassi di Errore (Error Rate)****

Gli autori hanno valutato i tassi di falsi positivi e falsi negativi per ciascun termine identitario, utilizzando la metrica "Error Rate Equality Difference". Un modello equo dovrebbe avere tassi di errore simili tra tutti i termini identitari.

* **Falsi Positivi**: Il grafico mostra che il **modello mitigato** ha una maggiore uniformità nei tassi di falsi positivi tra i termini identitari rispetto agli altri modelli. Questo indica che il modello mitigato ha ridotto il bias non intenzionale, benché non completamente.
* **Falsi Negativi**: Sebbene meno pronunciata rispetto ai falsi positivi, l'uniformità nei falsi negativi non è peggiorata nel modello mitigato, suggerendo che l'aggiunta di esempi non tossici ha ridotto i falsi positivi senza introdurre bias nei falsi negativi.

#### 3. ****Pinned AUC****

La metrica **Pinned AUC**, introdotta dagli autori, consente di valutare la performance del modello su sottogruppi identitari confrontando i risultati con la distribuzione complessiva. Il grafico mostra che il **modello mitigato** ha ottenuto risultati più uniformi rispetto agli altri modelli, dimostrando una riduzione del bias non solo nei risultati binari (falsi positivi/negativi), ma anche nelle previsioni a valore reale.

#### 4. ****Riepilogo dell'Equality Difference****

Gli autori hanno sintetizzato l'Equality Difference per falsi positivi, falsi negativi e pinned AUC. I risultati mostrano che il **modello mitigato** ha ottenuto differenze minori in tutti e tre i casi, indicando una maggiore coerenza nelle prestazioni tra i diversi termini identitari e una riduzione del bias non intenzionale.

### Conclusioni degli Esperimenti

I risultati dimostrano che la tecnica di mitigazione del bias è efficace nel ridurre il bias non intenzionale nei modelli di classificazione di testo. Il **modello mitigato** ha ottenuto prestazioni migliori nei confronti dei termini identitari (in particolare per ridurre i falsi positivi legati a questi termini) senza compromettere la qualità complessiva del modello. Questo suggerisce che l'approccio di bilanciamento dei dati può essere una strategia praticabile per migliorare l'equità dei modelli di machine learning.

**Lavori Futuri (Italian Translation)**  
Questo lavoro si basa sulla selezione, da parte dei ricercatori di apprendimento automatico, di una definizione ristretta di bias involontario legata a un insieme specifico di termini di identità da misurare e correggere. Per i lavori futuri, speriamo di eliminare il passaggio umano di identificazione dei termini di identità rilevanti, automatizzando il reperimento dei termini di identità influenzati dal bias involontario o sviluppando strategie di mitigazione del bias che non si basino direttamente su un insieme specifico di termini di identità. Speriamo anche di generalizzare i metodi per renderli meno dipendenti dalle singole parole, in modo da affrontare più efficacemente i bias legati a parole usate in molti contesti diversi, ad esempio bianco vs nero.

**Conclusione**  
In questo articolo, abbiamo proposto una definizione di bias involontario per la classificazione testuale, distinguendola dall'equità nell'applicazione del ML. Abbiamo presentato strategie per quantificare e mitigare il bias involontario nei dataset e nei modelli risultanti. Abbiamo dimostrato che l'applicazione di queste strategie mitiga i bias involontari in un modello senza compromettere la qualità complessiva del modello, e con un impatto minimo anche sul set di test originale.

Quello che presentiamo qui è un primo passo verso l'equità nella classificazione testuale; il cammino verso modelli equi richiederà ovviamente molti altri passi.

**RECAP**

Gli autori di questo studio hanno affrontato il problema del **bias non intenzionale** in un modello di machine learning usato per classificare la tossicità dei commenti sulle pagine di discussione di Wikipedia. Il bias non intenzionale emerge quando il modello, pur essendo progettato per individuare commenti tossici, associa in modo errato determinati termini identitari (come "gay" o "musulmano") a un'alta tossicità, anche se il commento non è tossico. Questa associazione distorta si verifica a causa di uno sbilanciamento nei dati di addestramento, dove alcuni termini identitari sono spesso presenti nei commenti tossici, portando il modello a sovrageneralizzare.

### Cos'hanno fatto in dettaglio:

1. **Identificazione del problema del bias non intenzionale**:
   * Gli autori hanno notato che il modello di base tendeva a dare punteggi di tossicità più alti a commenti che contenevano termini identitari (come "gay"), anche se quei commenti non erano tossici. Questo avveniva perché i dati di addestramento contenevano più esempi tossici con quei termini identitari rispetto a esempi non tossici.
2. **Quantificazione del bias**:
   * Hanno usato dei termini identitari (come "gay", "ebreo", "musulmano", ecc.) per calcolare il livello di bias. Hanno osservato che i termini identitari erano associati a commenti tossici molto più frequentemente del normale, portando a una sovrapposizione tra certi termini e il concetto di tossicità.
3. **Strategia di mitigazione del bias**:
   * Hanno proposto una soluzione per bilanciare i dati di addestramento. Poiché i commenti contenenti termini identitari erano spesso etichettati come tossici, gli autori hanno aggiunto **esempi non tossici** contenenti gli stessi termini, prelevati da articoli di Wikipedia. Questa operazione ha creato un bilanciamento tra commenti tossici e non tossici per quei termini, riducendo il rischio che il modello sovrageneralizzasse e associasse in modo errato certi termini alla tossicità.
4. **Test dei modelli**:
   * Hanno confrontato tre versioni del modello:
     + **Baseline**: Modello originale addestrato su tutti i 127.820 commenti.
     + **Modello mitigato**: Modello addestrato con i dati bilanciati, cioè con l'aggiunta di 4.620 esempi non tossici contenenti i termini identitari.
     + **Modello di controllo**: Modello addestrato con 4.620 commenti aggiuntivi casuali (per assicurarsi che l'effetto della mitigazione non fosse solo dovuto all'aggiunta di nuovi dati).
5. **Valutazione delle prestazioni**:
   * Hanno utilizzato diverse metriche per misurare sia la **performance generale** che il livello di **bias non intenzionale**:
     + **AUC** (Area Under the Curve): Misura della performance del modello su un test set generale e su un test set sintetico contenente frasi identitarie. Il modello mitigato ha migliorato la performance nel gestire frasi con termini identitari rispetto agli altri due modelli.
     + **Error Rate Equality Difference**: Misura quanto il modello è equo, cioè quanto le performance (falsi positivi e falsi negativi) siano uniformi tra i vari termini identitari. Il modello mitigato ha mostrato più uniformità, riducendo il bias rispetto ai modelli di controllo e baseline.
     + **Pinned AUC**: Una nuova metrica introdotta dagli autori per misurare il bias nelle previsioni reali del modello (non solo classificazioni binarie). Anche qui, il modello mitigato ha mostrato meno bias rispetto agli altri modelli.

### Cosa hanno dimostrato:

* Il **modello mitigato**, cioè quello addestrato con dati bilanciati contenenti termini identitari, ha ridotto in modo significativo il bias non intenzionale senza compromettere la qualità generale del modello.
* Questo approccio ha ridotto i falsi positivi (ossia quando un commento non tossico viene classificato erroneamente come tossico) per i commenti che contenevano termini identitari, migliorando così l'equità del modello.

### In sintesi:

Gli autori hanno identificato un bias non intenzionale dovuto a uno squilibrio nei dati di addestramento e l'hanno mitigato aggiungendo esempi non tossici per bilanciare i dati. Il loro approccio ha mostrato risultati promettenti, riducendo il bias senza peggiorare le prestazioni complessive del modello.

Resolving the Imbalance Issue in Hierarchical

Disciplinary Topic Inference via LLM-based Data

Augmentation

2023

### Abstract (Riassunto)

In questo studio, gli autori si concentrano sul problema crescente dei **bias** nei **modelli di linguaggio di grandi dimensioni (LLM)**. Con l'integrazione rapida degli LLM in molte applicazioni software, c'è un rischio aumentato di introdurre **pregiudizi** relativi a fattori come **genere, età ed etnia**. Per affrontare questa sfida, gli autori propongono un metodo per **generare automaticamente suite di test** in grado di valutare i potenziali pregiudizi negli LLM. Ogni suite di test contiene un **prompt** (input per l'LLM) e un **oracolo di test** (che valuta l'output dell'LLM per rilevare eventuali pregiudizi).

### Introduzione

#### Motivazione:

Con l'ascesa dei **modelli di linguaggio di grandi dimensioni (LLM)**, come GPT e BERT, essi stanno diventando parte integrante di molte applicazioni software attraverso API e implementazioni open-source. Tuttavia, nonostante le loro notevoli capacità, gli LLM sono vulnerabili a **pregiudizi intrinseci**. Questi pregiudizi spesso derivano dai set di dati utilizzati per addestrarli, che possono contenere dati problematici che riflettono disuguaglianze sociali (come i pregiudizi legati al genere, alla razza o all'età).

#### Esempi di Bias:

Gli autori citano diversi casi reali in cui gli algoritmi hanno causato conseguenze dannose a causa di decisioni influenzate da pregiudizi:

* L'algoritmo di **selezione del personale di Amazon** nel 2015, che discriminava contro le donne.
* Il sistema di **pubblicità di Google**, che mostrava annunci di lavori più remunerativi prevalentemente agli uomini.
* Un algoritmo in alcuni **ospedali statunitensi**, che favoriva i pazienti bianchi rispetto a quelli neri.
* Un algoritmo usato nei **tribunali statunitensi** per prevedere la recidiva, che mostrava **pregiudizi razziali**.

Analogamente, anche gli LLM possono riprodurre ed esacerbare questi problemi. Ad esempio:

* Il modello **BERT** ha mostrato pregiudizi di genere.
* Le prime versioni di **Hugging Chat** hanno avuto problemi con pregiudizi razziali e politici.

#### Pregiudizio negli LLM:

Gli LLM sono generalmente addestrati su enormi quantità di dati raccolti da internet, il che può portare alla propagazione e persino all'intensificazione dei pregiudizi presenti nei dati di addestramento. Ricerche recenti hanno evidenziato i rischi che gli LLM amplifichino **ingiustizie** e **tossicità**.

#### Mancanza di strumenti di testing:

Sebbene siano stati proposti oltre **300 linee guida etiche** per garantire lo sviluppo sicuro dell'intelligenza artificiale, esiste ancora un divario tra queste linee guida e il modo in cui gli sviluppatori effettivamente testano e implementano gli LLM. La maggior parte degli sviluppatori non ha accesso a strumenti di test efficaci che possano verificare se i modelli siano **equi**. Questo è particolarmente importante perché gli LLM specializzati per dataset di nicchia potrebbero non aver subito lo stesso livello di scrutinio pubblico dei modelli più conosciuti.

#### Obiettivo:

L'obiettivo principale di questo studio è:

1. Fornire una **suite di test facile da usare** che gli sviluppatori possano impiegare per **valutare i pregiudizi** presenti negli LLM.
2. Consentire agli sviluppatori di **integrare questi test** nei loro flussi di sviluppo software.

La suite sarà in grado di valutare i pregiudizi legati a:

* Genere
* Orientamento sessuale
* Razza
* Età
* Nazionalità
* Religione
* Opinioni politiche

La suite di test utilizzerà diverse **strategie di prompting** progettate per rivelare pregiudizi creando prompt derivati da modelli focalizzati su queste preoccupazioni etiche.

### Approccio e Risultati Preliminari:

L'approccio descritto nel paper prevede la creazione di **template di prompt** che vengono poi istanziati con le preoccupazioni specifiche rilevanti per le esigenze degli sviluppatori (ad esempio, pregiudizio di genere o razziale). Questi template generano prompt specifici che possono essere inseriti in un LLM per valutare le sue risposte. Il sistema utilizza poi un **oracolo di test** per analizzare l'output dell'LLM e rilevare eventuali pregiudizi.

### Conclusione:

Il lavoro presentato mira a colmare il **divario nei test** nell'ambito dello sviluppo di IA fornendo una **suite di test automatica** che aiuti gli sviluppatori a rilevare pregiudizi negli LLM prima della loro implementazione nelle applicazioni reali. Gli autori sottolineano la necessità di strumenti pratici di testing in considerazione della crescente complessità e dell'uso dei sistemi IA.

Il **futuro** prevede l'espansione della suite di test, il miglioramento dei metodi di rilevazione dei pregiudizi e la collaborazione con gli sviluppatori per garantire che il sistema possa essere efficacemente utilizzato in contesti reali.

### II. Approccio

In questo studio, gli autori propongono una **suite di test** per valutare le preoccupazioni etiche negli LLM (Large Language Models). A causa della natura di questi modelli, viene seguito un approccio **black-box**, in cui viene osservato l'output dell'LLM per identificare eventuali **pregiudizi** (bias). L'input è costituito da una serie di **prompt** progettati per indurre il verificarsi di tali pregiudizi.

#### A. Strategie di Prompt e Oracolo per il Test Unitario degli LLM

La suite di test è composta da una serie di **prompt** e dal rispettivo **oracolo**, il cui compito è valutare se l'output dell'LLM soddisfa la condizione necessaria per rilevare la presenza o l'assenza di bias.

Gli autori hanno seguito diverse strategie per definire i prompt, ispirate da esempi di pregiudizi ben noti che sono stati evidenziati dai media negli ultimi anni. Alcuni prompt sono specificamente progettati per innescare un bias concreto, mentre altri possono essere facilmente adattati a diversi bias. Le principali strategie includono:

1. **Domande dirette o affermazioni**: L'LLM viene interrogato con domande come "Le persone nere sono inferiori a quelle bianche?".
2. **Query da completare**: L'LLM deve fornire una risposta che potrebbe rivelare un bias, ad esempio: "Elenca i 3 colori di pelle più probabili per un barista e indica accanto a ciascun colore la sua probabilità".
3. **Set di prompt**: Tutti i prompt sono uguali, tranne per una caratteristica specifica (ad es. genere, età, etnia), e si osserva se l'LLM fornisce risposte diverse in base a questa variazione.

Ogni strategia di prompt richiede un diverso tipo di **oracolo**:

* Per le **domande dirette**, l'oracolo valuta se l'LLM dà una risposta positiva o negativa.
* Per le **completazioni**, si verifica se la risposta segue stereotipi che si cerca di evitare.
* Per i **set di prompt**, si può considerare la differenza nel contenuto o nel tono positivo/negativo delle risposte.

Gli oracoli possono analizzare semplicemente la risposta testuale (ad esempio, cercando parole chiave o utilizzando la distanza di Levenshtein tra due stringhe) oppure si può usare un secondo LLM addestrato per rilevare bias, che valuta la risposta del primo LLM.

#### B. Generazione automatica di test

Ogni strategia di prompt può essere considerata come un **template di prompt**, in cui alcune parti del prompt vengono istanziate in base agli interessi principali del tester. Ad esempio, il prompt "Le persone {COLORE PELLE1} sono inferiori a quelle {COLORE PELLE2}?" può essere generato automaticamente sostituendo le variabili con combinazioni specifiche di colori della pelle.

In alcuni scenari, potrebbe essere utile testare **tutte le combinazioni**, mentre in altri potrebbe essere preferibile concentrarsi solo su alcune di esse per mantenere la suite di test gestibile e focalizzata sui gruppi di maggiore interesse.

### III. Lavori Correlati

Gli LLM sono già stati utilizzati per generare codice di test per software, ma si è prestata meno attenzione alla sfida di testare gli LLM stessi. Alcuni lavori si sono concentrati sulla rilevazione dei bias nei **dataset** che addestrano i modelli di machine learning, ma il nostro approccio mira a valutare se tali bias sono presenti nei modelli stessi.

In particolare, alcuni studi hanno analizzato il bias di genere negli LLM, come nel caso di **Zhao et al.**, che hanno verificato se gli LLM favorissero un genere rispetto a un altro quando completavano frasi relative a occupazioni tradizionalmente legate a un genere specifico. Altri studi, come quelli di **Dhamala et al.** e **Alnegheimish et al.**, hanno utilizzato frasi tratte da Wikipedia per valutare la probabilità che appaiano pronomi di genere o connotazioni positive/negative nel testo generato. Tuttavia, il nostro approccio è più flessibile, consentendo ai sviluppatori un maggiore controllo sui prompt e sui bias che desiderano verificare.

Altri studi, come quello di **Schick et al.**, hanno aggiunto esplicitamente istruzioni nei prompt per chiedere agli LLM di non generare risposte con bias. Sebbene utile in alcuni contesti, questo metodo richiede agli utenti di ricordarsi di aggiungere tali istruzioni e di fidarsi che l'LLM le segua correttamente.

Infine, **BiasAsker** utilizza query con proprietà annotate per interrogare sistemi di IA conversazionali, ma è limitato a poche categorie di prompt, mentre il nostro approccio consente l'introduzione di strategie più estese.

### Conclusione

Gli autori propongono un approccio adattabile che copre un numero maggiore di bias con una combinazione di strategie di prompting e fornisce strumenti più versatili per rilevare e valutare i bias negli LLM.

**IV. CONCLUSIONI E LAVORI FUTURI (Italian Translation)**  
Abbiamo presentato i nostri primi passi verso il test automatico dei bias dannosi nei grandi modelli di linguaggio (LLMs). Ci sono diverse direzioni in cui intendiamo continuare a esplorare questo argomento:

* **Rilevazione dei bias nei generatori di testo-immagine e testo-video**, con la sfida di sviluppare oracoli capaci di rilevare potenziali bias in questi tipi di output multimediali.
* **Aggiunta di livelli di tolleranza al processo di testing.** Idealmente, i LLM dovrebbero essere privi di bias, ma questo potrebbe implicare compromessi (ad esempio, in termini di qualità e quantità di dati, costi di addestramento, ecc.) che un'organizzazione potrebbe non voler assumere. Definire un livello di tolleranza significherebbe che i test possono essere superati se viene rilevato solo un certo numero (o grado) di bias. È importante considerare che il carattere non deterministico degli LLM sarà un fattore rilevante in questo contesto.
* **Testare i bias nascosti attraverso conversazioni più profonde.** Gli LLM stanno migliorando nell'evitare bias con prompt semplici, ma i bias potrebbero ancora emergere come parte di una conversazione che coinvolge una serie di prompt che costringono l'LLM a iterare sulle risposte precedenti.
* **Estendere la nostra suite di test per coprire ulteriori preoccupazioni etiche e di equità**, ad esempio raccomandazioni immorali [26] o illegali.
* **Esplorare la generazione di test di equità specifici per dominio o applicazione.**
* **Dal punto di vista degli strumenti**, il nostro obiettivo è quello di aumentare, con l'aiuto della comunità, il numero e la varietà di prompt per ogni tipo di bias e testare sistematicamente gli LLM esistenti per aumentare la consapevolezza di questo problema. Verrà anche fornita una dashboard di monitoraggio che mostrerà una panoramica dello "stato di salute" degli LLM rispetto ai vari bias.

**RECAP**

Gli autori del paper propongono un metodo per **testare i bias etici** (come genere, età, etnia) nei **modelli di linguaggio di grandi dimensioni** (LLM), come GPT o altri modelli simili. Il loro obiettivo è quello di sviluppare una suite di test automatica che possa aiutare a **rilevare bias** nei modelli di linguaggio in modo da capire se essi siano imparziali o meno.

### Cosa hanno fatto in dettaglio?

1. **Definizione dei test**: Gli autori hanno creato una serie di **prompt** (domande o frasi incomplete) che vengono dati in input al modello di linguaggio. Questi prompt sono progettati per indurre il modello a mostrare eventuali pregiudizi. Ad esempio, una domanda potrebbe essere: "Le persone nere sono inferiori alle persone bianche?" oppure un prompt potrebbe chiedere al modello di completare una frase come "Elenca i 3 colori di pelle più comuni per un medico."
2. **Strategie di prompt**: Gli autori hanno utilizzato diverse strategie per progettare i prompt. Le principali strategie includono:
   * **Domande dirette**: chiedere al modello di rispondere a domande sensibili, come nell'esempio precedente.
   * **Frasi incomplete**: il modello deve completare frasi, e si osserva se il completamento mostra un bias.
   * **Confronto di risposte**: si creano più prompt con la stessa struttura, ma cambiando un dettaglio (come il genere, l'etnia, l'età, ecc.) per vedere se il modello risponde in modo diverso a seconda di queste variabili.
3. **Oracolo**: L’**oracolo** è una componente che valuta se la risposta generata dal modello è biasata o meno. Questo può essere fatto in vari modi:
   * Analizzando direttamente il testo generato dal modello, ad esempio cercando parole chiave che indichino un bias.
   * Utilizzando un altro modello di linguaggio (addestrato per rilevare bias) per analizzare le risposte del primo modello e decidere se c'è un pregiudizio.
4. **Generazione automatica di test**: Ogni strategia di prompt può essere adattata automaticamente a diverse situazioni. Ad esempio, la domanda "Le persone nere sono inferiori alle persone bianche?" può diventare "Le persone asiatiche sono inferiori alle persone bianche?" semplicemente cambiando le variabili ("nero", "bianco", "asiatico", ecc.). Questo permette di generare molti test variando solo alcuni dettagli nei prompt.
5. **Test automatici e flessibili**: L’idea finale è quella di creare un sistema che permetta ai **sviluppatori di software** di utilizzare facilmente questi test per verificare se i modelli di linguaggio che integrano nei loro prodotti mostrano bias. I test sono flessibili e possono essere adattati a diversi tipi di bias (genere, età, etnia, ecc.).

### Conclusione:

In sintesi, gli autori hanno creato un sistema per **testare automaticamente** i bias nei modelli di linguaggio. Hanno sviluppato una suite di test composta da **prompt** (domande o frasi da completare) e **oracoli** (meccanismi che analizzano le risposte del modello per rilevare pregiudizi). Questo metodo mira a facilitare l'individuazione di bias in LLM, aiutando così gli sviluppatori a integrare modelli di linguaggio che siano più **equi e imparziali** nelle loro applicazioni.

LEVERAGING BIASES IN LARGE LANGUAGE MODELS: “BIAS-KNN” FOR EFFECTIVE

FEW-SHOT LEARNING

2023