**Stato dell’arte**

Negli ultimi anni, la mitigazione dei bias nei modelli di linguaggio è diventata un tema cruciale nel campo del machine learning, con numerosi studi che hanno affrontato il problema. Tuttavia, gran parte delle ricerche si è concentrata su bias specifici, come quello di genere o razza, utilizzando tecniche limitate a contesti predefiniti e spesso basate su frasi artificiali. Questi approcci, sebbene utili, non riescono a cogliere la complessità dei bias latenti che emergono nel linguaggio naturale e in contesti più ampi, come etnia, cultura o background socioeconomico.

Il mio framework evolve lo stato dell'arte integrando **Knowledge Graphs (KGs)** per arricchire i dati con relazioni semantiche tra concetti, migliorando la capacità di rilevare bias complessi. Utilizza una **rappresentazione multivettoriale** delle frasi, che combina embedding e informazioni semantiche derivate dai KGs, offrendo una visione più ricca e articolata delle entità e delle loro interazioni. Inoltre, il framework applica tecniche di **clusterizzazione semantica**, che vanno oltre la somiglianza testuale, permettendo di identificare bias latenti tra frasi semanticamente correlate.

Un'altra innovazione chiave è l'uso di **modelli generativi** per creare scenari estremi e testare la resilienza del modello in situazioni che potrebbero attivare bias nascosti. Questo approccio non solo migliora la rilevazione dei bias, ma garantisce una maggiore equità e trasparenza nelle decisioni dei modelli di linguaggio. Nel complesso, il mio framework rappresenta un'evoluzione significativa rispetto agli approcci tradizionali, poiché offre una soluzione più robusta, completa e applicabile a una gamma più ampia di contesti e scenari reali.

### Stato dell'Arte esteso

La mitigazione del bias nei modelli di linguaggio è una delle sfide più rilevanti e urgenti nell'attuale panorama del machine learning. Con l'aumento dell'uso di modelli di linguaggio di grandi dimensioni (LLM), come BERT, GPT e RoBERTa, in applicazioni che influenzano decisioni cruciali nella vita quotidiana, il rischio che questi modelli perpetuino o amplifichino bias sociali è diventato un tema centrale. L'obiettivo di questo stato dell'arte è esaminare i principali lavori in questo campo, individuandone i limiti e illustrando come il framework proposto rappresenti un'evoluzione rispetto a tali approcci.

#### Lavori di Riferimento

Uno dei primi studi rilevanti è **“Towards Understanding and Mitigating Social Biases in Language Models” (2021)**, che ha esplorato il problema dei bias nei modelli pre-addestrati come BERT. Questo lavoro ha evidenziato come i modelli possano riflettere pregiudizi presenti nei dati di addestramento, proponendo soluzioni di debiasing attraverso la manipolazione delle rappresentazioni latenti. Tuttavia, il focus ristretto a bias generici, come il genere e la razza, rappresenta un limite significativo, poiché non viene considerata la complessità delle relazioni semantiche tra entità culturali ed etniche.

In **“Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach” (2023)**, gli autori hanno utilizzato una rete avversaria per ridurre il bias di genere nei modelli transformer. Sebbene questo approccio abbia mostrato un miglioramento nell'equità, il focus esclusivo sul genere e l'assenza di tecniche per considerare altre dimensioni del bias, come etnia e contesto culturale, ne limitano la generalizzabilità. Inoltre, non viene sfruttata la ricchezza delle relazioni semantiche offerte dai **Knowledge Graphs (KGs)**, un elemento che potrebbe arricchire l'efficacia delle tecniche di mitigazione.

**“StereoSet: Measuring Stereotypical Bias in Pretrained Language Models” (2021)** ha introdotto un dataset per misurare bias stereotipici nei modelli di linguaggio pre-addestrati. Sebbene questo studio abbia contribuito a quantificare bias legati al genere, alla razza e alla religione, l'uso di frasi artificiali limita la capacità di catturare la complessità del linguaggio naturale. La mancanza di integrazione con i KGs impedisce inoltre di identificare bias impliciti basati su relazioni più complesse tra concetti.

Un'altra proposta interessante è **“Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal” (2022)**, dove viene introdotta l'inversione dei ruoli professionali nei dati di addestramento per ridurre il bias di genere. Questo approccio ha mostrato buoni risultati nel mitigare il bias nei ruoli professionali, ma resta limitato all'ambito di applicazione specifico e non affronta bias legati ad altre dimensioni, come la cultura e il background etnico.

Infine, **“BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems” (2022)** ha introdotto tecniche di testing metamorfico per rilevare bias nei sistemi di sentiment analysis. Tuttavia, il lavoro si concentra su un numero limitato di bias (genere e occupazione) e si basa su template predefiniti, che non catturano scenari più complessi o realistici.

#### Limiti dei Lavori Esistenti

I limiti principali dei lavori esistenti includono:

* **Focus limitato su specifici bias**: La maggior parte degli studi si concentra su bias di genere o di occupazione, trascurando altre dimensioni come etnia, cultura e background socioeconomico.
* **Dipendenza da frasi artificiali**: Molti approcci, come StereoSet, si basano su frasi generate artificialmente, riducendo la capacità di rilevare bias che emergono in contesti reali.
* **Assenza di KGs**: L'assenza di Knowledge Graphs, che possono fornire relazioni semantiche tra concetti e contesti, limita la profondità dell'analisi e la capacità di rilevare bias impliciti.

#### Evoluzione del Framework Proposto

Il framework proposto affronta questi limiti con una serie di innovazioni che espandono l'ambito di rilevazione e mitigazione del bias nei modelli di linguaggio.

1. **Raccolta e Incrocio di Dataset Diversificati**: A differenza dei lavori esistenti che si basano su dati limitati o artificiali, il framework raccoglie dataset da fonti eterogenee, come social media, recensioni e notizie. L'integrazione dei **Knowledge Graphs (KGs)** arricchisce i dati con relazioni semantiche tra entità, migliorando la rilevazione di bias culturali, etnici e storici.
2. **Rappresentazione Multivettoriale con KGs**: Il framework propone una rappresentazione multivettoriale delle frasi che combina embedding generati da modelli di linguaggio con informazioni semantiche estratte dai KGs. Questa rappresentazione arricchita aumenta la capacità di rilevare bias latenti che altrimenti sfuggirebbero ai modelli basati solo su rappresentazioni testuali.
3. **Clusterizzazione Semantica**: Il framework implementa tecniche di clusterizzazione basate sulle relazioni semantiche tra concetti, non solo sulla somiglianza testuale. Questo approccio permette di identificare bias latenti anche in frasi che potrebbero apparire distanti dal punto di vista lessicale, ma che condividono connessioni semantiche.
4. **Tecniche di Disentanglement e Regolarizzazione**: Uno degli aspetti innovativi del framework è l'integrazione di tecniche di disentanglement, che separano le rappresentazioni degli attributi sensibili (come genere o etnia) dalle altre caratteristiche latenti, riducendo il rischio di bias impliciti nelle predizioni.
5. **Modelli Generativi per Stress Testing**: Un'ulteriore evoluzione rispetto ai lavori esistenti è l'uso di modelli generativi per creare scenari estremi e testare la resilienza del modello. Questo approccio permette di valutare come i modelli rispondono a input che potrebbero attivare bias latenti, offrendo una capacità di testing molto più robusta.

Un confronto interessante può essere fatto con il pacchetto R **fairmodels**, descritto nel lavoro **"Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models" (2022)**. Questo strumento mira a identificare e mitigare il bias nei modelli di classificazione binaria. Sebbene offra un'interfaccia versatile per la diagnosi di bias e l'equità dei modelli, si limita al dominio della classificazione binaria e non considera le potenzialità dei KGs o l'integrazione di rappresentazioni semantiche più ricche come quelle proposte nel framework. Inoltre, il framework proposto supera **fairmodels** grazie a un approccio più completo alla mitigazione del bias, che include l'uso di clusterizzazione semantica e tecniche avanzate di interpretabilità.

#### Conclusione

In sintesi, il framework proposto rappresenta un passo avanti significativo rispetto ai lavori esistenti, estendendo la capacità di rilevamento e mitigazione dei bias attraverso l'uso di dataset diversificati, l'integrazione di Knowledge Graphs e tecniche avanzate di interpretabilità e clusterizzazione semantica. Questo approccio non solo affronta il problema del bias in modo più completo, ma permette anche di gestire scenari complessi e dinamici, migliorando l'equità e la trasparenza nei modelli di linguaggio.

**PAPERS DI CONFRONTO da confrontare o utili nell’implementazione del framework**

CATEGORIA DETECTION:

* Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach (2023)
* Mitigating Gender Bias in Distilled Language Modelsvia Counterfactual Role Reversal (2022)
* Fairway: AWay to Build Fair ML Software (2020) [Meno buono]
* Measuring and Mitigating Unintended Bias in Text Classification (2019) [Meno buono]
* Identifying Gender Bias in Online Crime News Indonesia Using Word Embedding (2024) [Meno buono]
* Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva (2020)
* Detection of Latent Gender Biases in Data and Models Using the Approximate Generalized Inverse Method (2024)
* Context-Aware Deep Markov Random Fields for Fake News Detection (2021) [Meno buono]

CATEGORIA MITIGATION (DETECTION INCLUSA):

* Fine-Tuning Language Models to Mitigate Gender Bias in Sentence Encoders (2022)
* COMPENSATORY DEBIASING FOR GENDER IMBALANCES IN LANGUAGE MODELS (2023)
* Ensuring Transparency and Fairness in AI Decision-Making Processes Influenced by large language Models (2024)
* Examining and mitigating gender bias in text emotion detection task (2022)
* fairmodels: a Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models (2022)
* On Evaluating and Mitigating Gender Biases in Multilingual Settings (2023)

CATEGORIA TESTING:

* BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems (2022)
* Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-based DataAugmentation (2023)

CATEGORIA BIAS CONTESTUALE:

* LEVERAGING BIASES IN LARGE LANGUAGE MODELS: “BIAS-KNN” FOR EFFECTIVE FEW-SHOT LEARNING (2023)
* Directional Pairwise Class Confusion Bias and Its Mitigation (2022)

CATEGORIA BENCHMARK:

* CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models (2020)
* The Evolution of Large Language Model: Models, Applications and Challenges (2024)
* Exploring Bias Evaluation Techniques for Quantifying Large Language Model Biases (2023)
* BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation (2021)
* StereoSet: Measuring stereotypical bias in pretrained language models (2021)
* Towards Understanding and Mitigating Social Biases in Language Models (2021)

**FRAMEWORK NEL DETTAGLIO**

**MODULO DI DETECTION**

**Step 1: Raccolta e Incrocio di Dataset Testuali Diversificati**

**Descrizione**: In questa fase, raccogli dataset da diverse fonti (enciclopediche, social media, recensioni, notizie) con attenzione a riflettere una vasta gamma di contesti culturali, sociali, etnici e di genere. Integra i **Knowledge Graphs (KGs)** come fonte di arricchimento per migliorare il contesto delle entità e delle relazioni tra concetti. Poiché è necessario un confronto quantitativo con altri lavori sarebbe utile usare dataset noti come StereoSet, CrowS-Pairs, WinoBias ecc.

**Come utilizzare i KGs**: Utilizza KGs come **Wikidata**, **DBpedia** o **Freebase** per migliorare i tuoi dataset. Ad esempio, se stai raccogliendo dati di recensioni di prodotti, puoi estrarre informazioni strutturate da DBpedia che descrivono i prodotti stessi, aggiungendo attributi e relazioni come la categoria del prodotto, il produttore, ecc.

**Esempio**:

* Supponiamo di raccogliere un dataset di frasi relative a diverse professioni. Con l’aiuto di un KG come Wikidata, puoi arricchire il dataset aggiungendo dettagli come il **genere predominante** in quella professione, la **storia culturale** della professione e altre informazioni pertinenti.
* Una frase come "Il dottore è arrivato in ritardo" può essere arricchita con informazioni da Wikidata che indicano se la professione medica è storicamente dominata da uomini o da donne in certi contesti geografici. Questo aiuta a riconoscere bias impliciti o espliciti legati al genere.

**Dataset noti**

### 1. ****Bias di Genere****

* **Winogender Schemas**: Progettato per misurare il bias di genere nei modelli di linguaggio. Si basa su frasi ambigue (schema Winograd) dove il genere del pronome viene influenzato da stereotipi impliciti.
* **WinoBias**: Simile a Winogender, ma focalizzato sul bias di genere in contesti più complessi. Utile per testare come i modelli associano ruoli professionali a determinati generi.
* **Gender Bias in Context (GBC)**: Un dataset che include frasi più complesse per valutare il bias di genere in diversi contesti. Offre esempi più realistici e non artificiali.

### 2. ****Bias Etnico e Razziale****

* **StereoSet**: Misura il bias stereotipico non solo su genere ma anche su razza e religione. È utile per testare più dimensioni di bias contemporaneamente. Contiene frasi generate per catturare stereotipi e frasi neutrali per confronto.
* **Bias in Bios**: Un dataset che contiene biografie professionali e che può essere utilizzato per misurare il bias di genere e razza nelle professioni. È utile per valutare l'accuratezza dei modelli nell'evitare bias in contesti di occupazione.
* **Ethnic Bias Dataset**: Un dataset specifico per la rilevazione di bias etnico. Potresti cercare dataset simili o costruire dataset ad hoc utilizzando fonti eterogenee, come notizie e social media, per analizzare bias legati a etnia e cultura.

### 3. ****Bias Generale nei Modelli di Linguaggio****

* **CrowS-Pairs**: Un dataset creato per misurare bias su molteplici dimensioni (genere, razza, orientamento sessuale, religione, disabilità, ecc.). Le frasi sono progettate per avere una variante stereotipica e una variante non stereotipica, rendendolo utile per valutare diverse forme di bias contemporaneamente.
* **Toxic Comments Dataset (Jigsaw)**: Un dataset di commenti tossici online, utile per analizzare bias in frasi che contengono linguaggio offensivo o discriminatorio. Ottimo per valutare bias nei modelli che devono fare moderazione di contenuti.
* **The Real Toxicity Prompts Dataset**: Un altro dataset utile per valutare il bias nei modelli generativi. Contiene prompt che possono scatenare risposte tossiche da modelli di linguaggio come GPT, ed è utile per stress testing.

### 4. ****Bias nelle Professioni e Occupazioni****

* **Occupation Classification Dataset**: Un dataset che include frasi su varie occupazioni, utile per valutare come i modelli associano determinati ruoli a certi generi o gruppi etnici. Viene spesso usato per studiare il bias nei ruoli professionali.
* **Fairface Dataset**: Un dataset di immagini di volti etichettati per genere, etnia, e altre caratteristiche. Sebbene si tratti di immagini, potrebbe essere utile nel contesto della tua ricerca, ad esempio, per studiare la generazione di descrizioni linguistiche biasate a partire da dati visivi.

### 5. ****Dataset da Social Media****

* **Twitter Bias Dataset**: Contiene tweet raccolti per analizzare bias su vari aspetti, come razza, genere e orientamento politico. È utile per studiare bias in linguaggio più informale e spontaneo.
* **Tweep Bias**: Un altro dataset simile, focalizzato su come i modelli rispondono ai tweet, in particolare su questioni di bias etnico, religioso e di genere.

### 6. ****Dataset da Notizie e Recensioni****

* **NewsQA**: Un dataset di domande e risposte generate da articoli di notizie. È utile per studiare bias nel linguaggio dei media, specialmente quando si analizzano articoli su argomenti delicati come immigrazione, razza e genere.
* **Amazon Product Reviews**: Questo dataset è ricco di recensioni di prodotti che possono essere utilizzate per esaminare bias in diversi contesti (ad esempio, preferenze di genere per certi prodotti). Può essere arricchito con KGs per comprendere meglio come i prodotti vengono percepiti in base al genere o alla cultura.

### 7. ****Dataset Multiculturale e Multilingua****

* **XED (Cross-lingual Ethnic Diversity)**: Questo dataset fornisce esempi di linguaggio che trattano temi di diversità etnica in diversi contesti culturali e linguistici. Può essere utile per valutare bias nei modelli multilingua.
* **Multilingual Wikipedia Bias Dataset**: Un dataset che esplora il bias nei contenuti di Wikipedia in varie lingue, utile per analizzare come i modelli gestiscono il bias legato a cultura ed etnia in contesti globali.

**Abbinamento con KGs**

### 1. ****Bias di Genere****

* **Dataset**: Winogender Schemas, WinoBias, Gender Bias in Context (GBC)
* **KG suggerito**: **Wikidata**
  + **Motivazione**: Wikidata contiene relazioni dettagliate su persone, professioni e ruoli sociali che possono essere usate per arricchire frasi relative a bias di genere. Per esempio, puoi collegare una frase a concetti come "genere", "ruoli di genere" o "stereotipi di genere" e analizzare come questi concetti sono interconnessi in contesti storici e geografici.
  + **Esempio**: Se una frase nel dataset riguarda una professione (ad esempio, "Il dottore ha eseguito l'operazione"), puoi usare Wikidata per estrarre informazioni sul genere predominante nelle professioni mediche in vari contesti culturali e temporali, arricchendo così l'analisi del bias di genere.

### 2. ****Bias Etnico e Razziale****

* **Dataset**: StereoSet, Bias in Bios, Ethnic Bias Dataset
* **KG suggerito**: **DBpedia** o **YAGO**
  + **Motivazione**: Entrambi i KGs offrono informazioni ben strutturate su figure storiche, gruppi etnici e contesti culturali che possono essere utilizzati per esaminare stereotipi o pregiudizi etnici. DBpedia è particolarmente ricco in termini di dati estratti da Wikipedia, mentre YAGO può essere utile per dati etnici e di background culturale.
  + **Esempio**: In un dataset come StereoSet, una frase che menziona un gruppo etnico o una nazionalità (ad esempio, "Gli asiatici sono bravi in matematica") può essere arricchita con dati estratti da DBpedia o YAGO che riguardano la rappresentazione di questo gruppo in ambiti come l'istruzione, la storia o la cultura. Questo aiuta a individuare e analizzare il bias in un contesto più ampio.

### 3. ****Bias su Religione e Contesto Culturale****

* **Dataset**: CrowS-Pairs, Twitter Bias Dataset
* **KG suggerito**: **ConceptNet** o **WordNet**
  + **Motivazione**: ConceptNet è un KG che fornisce una rete semantica con relazioni tra concetti legati alla cultura, alla religione e agli stereotipi sociali. Puoi utilizzare le relazioni concettuali tra entità per comprendere meglio il contesto culturale e religioso dietro certe frasi. WordNet è particolarmente utile per relazioni semantiche su sinonimi, antonimi e gerarchie concettuali.
  + **Esempio**: Supponiamo che una frase come "I musulmani sono spesso collegati al terrorismo" appaia in un dataset. Usando ConceptNet, puoi individuare relazioni che collegano "musulmani" a concetti più ampi come "religione", "comunità", "conflitto" o "pregiudizio", il che ti permette di evidenziare il contesto culturale in cui il bias si manifesta e migliorare la mitigazione.

### 4. ****Bias nelle Professioni e Occupazioni****

* **Dataset**: Bias in Bios, Occupation Classification Dataset
* **KG suggerito**: **Wikidata** o **DBpedia**
  + **Motivazione**: Wikidata e DBpedia contengono informazioni ricche sulle professioni e sulle occupazioni, che includono dati storici, demografici e culturali. Questi grafi possono aiutarti a contestualizzare frasi relative ai bias di genere e razza nelle professioni, integrando relazioni tra professioni, stereotipi di genere e trend occupazionali.
  + **Esempio**: Se il dataset contiene frasi come "L'ingegnere ha progettato il ponte", puoi usare DBpedia per identificare dati storici sulla prevalenza di ingegneri di genere maschile o femminile e ampliare il contesto delle frasi, migliorando la capacità di rilevare bias nascosti.

### 5. ****Bias nei Media e Notizie****

* **Dataset**: NewsQA, Multilingual Wikipedia Bias Dataset
* **KG suggerito**: **Wikidata** o **Media Bias/Fact Check Knowledge Graph (MBFC)** (se disponibile)
  + **Motivazione**: Wikidata può essere utilizzato per collegare entità nei testi delle notizie, come figure politiche, eventi storici e istituzioni, e arricchire la comprensione semantica delle frasi. Il Media Bias/Fact Check Graph, se disponibile, può aiutarti a classificare notizie in base all'inclinazione politica e al contesto, migliorando l'analisi del bias nei media.
  + **Esempio**: Se stai analizzando articoli su immigrazione o diritti umani, puoi utilizzare Wikidata per collegare concetti come "immigrazione", "leggi sull'immigrazione", "diritti civili", e analizzare come questi concetti sono rappresentati nei media e nelle notizie.

### 6. ****Bias nei Social Media****

* **Dataset**: Twitter Bias Dataset, Tweep Bias
* **KG suggerito**: **OpenStreetMap** (per contesto geografico) + **ConceptNet** (per contesto semantico)
  + **Motivazione**: In contesti come Twitter, dove le frasi sono brevi e spesso decontestualizzate, OpenStreetMap può aiutare a fornire informazioni geografiche sulle entità menzionate nei tweet, mentre ConceptNet può arricchire il contesto semantico e culturale delle conversazioni. Questo aiuta a identificare bias regionali o legati a contesti specifici.
  + **Esempio**: Un tweet che menziona un gruppo etnico in una certa regione geografica (ad esempio, "Gli africani in Europa") può essere arricchito con dati di OpenStreetMap per identificare la provenienza geografica e con ConceptNet per contestualizzare concetti legati a "migrazione", "integrazione" e "pregiudizio".

### 7. ****Bias Multilingue****

* **Dataset**: XED (Cross-lingual Ethnic Diversity), Multilingual Wikipedia Bias Dataset
* **KG suggerito**: **BabelNet** o **DBnary**
  + **Motivazione**: Per analizzare bias in contesti multilingua, BabelNet e DBnary forniscono risorse lessicali e semantiche in più lingue. BabelNet, in particolare, integra WordNet e Wikipedia e può fornire relazioni multilingue tra concetti, utili per individuare bias culturali e etnici in diverse lingue.
  + **Esempio**: Se una frase in inglese o spagnolo menziona gruppi etnici o nazionali, puoi usare BabelNet per trovare corrispondenze concettuali in altre lingue e analizzare se il bias persiste tra lingue diverse.

**IDEA**

Normalizzare **StereoSet**, **CrowS-Pairs**, e **WinoBias** all'interno di un'unica metrica o framework è un approccio ambizioso e interessante. La sfida è creare una metrica unificata che possa misurare in modo coerente i diversi aspetti del bias trattati da ciascun dataset, mantenendo al contempo la varietà dei task e delle categorie di bias. Ecco un possibile approccio per creare questo framework:

**1. Definizione delle Componenti Chiave del Framework**

Prima di definire una metrica unificata, è importante identificare le componenti chiave che ogni benchmark misura:

* **StereoSet**: Misura sia **stereotypicality** (la tendenza del modello a completare le frasi in modo stereotipato) sia **accuracy** (la correttezza delle risposte non biasate).
* **CrowS-Pairs**: Si concentra sulla **preferenza** tra una frase bias e una frase neutra o inversa, misurando quindi quanto il modello preferisce una costruzione stereotipata.
* **WinoBias**: Testa la capacità del modello di evitare il **bias di genere** nella risoluzione della co-referenza (associando correttamente i pronomi ai soggetti senza stereotipi).

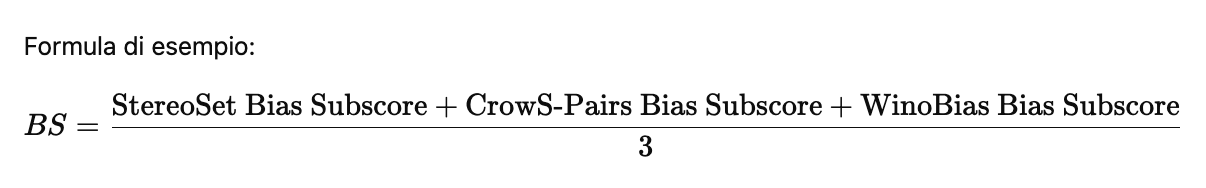
**2. Metrica Unificata Proposta**

Una metrica unificata potrebbe essere composta da più dimensioni per catturare sia il **livello di bias** sia la **precisione linguistica** del modello. Propongo di suddividere la metrica in due parti principali:

**2.1 Bias Score (BS):**

Misura il livello di bias stereotipato espresso dal modello. Questo punteggio sarà composto dalle valutazioni dei vari benchmark sui **bias di genere, razza, religione, orientamento sessuale, disabilità** ecc. Si potrebbero combinare i risultati da StereoSet, CrowS-Pairs e WinoBias in un'unica formula.

* **StereoSet Bias Subscore**: Derivato dalla "stereotypicality" score di StereoSet, che indica quanto il modello tende a preferire completamenti stereotipati.
* **CrowS-Pairs Bias Subscore**: Basato sulla percentuale di preferenze del modello per frasi bias rispetto a frasi non bias.
* **WinoBias Bias Subscore**: Basato sull'accuratezza nel collegare correttamente pronomi e soggetti senza cadere in stereotipi di genere.

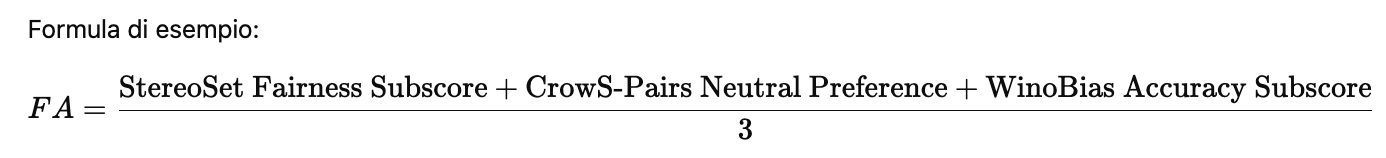
****

Questa formula aggrega i risultati dei tre benchmark, bilanciandoli equamente per ottenere un punteggio generale di bias. Se uno dei tre benchmark ha un focus su un tipo specifico di bias (es. genere), si può ponderare il punteggio in base alla categoria di bias, oppure normalizzare i risultati in modo che ogni tipo di bias pesi equamente.

#### 2.2 **Fairness-Accuracy Score (FA)**:

Questa parte della metrica misura quanto il modello è **accurato** nel completare i task linguistici, ma tenendo conto del fatto che evitare i bias deve essere prioritario. Sarà basata su due componenti:

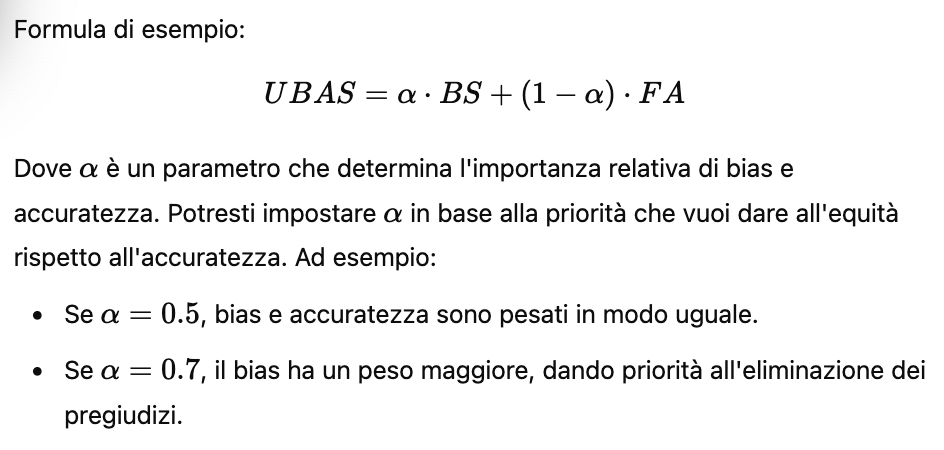
* **StereoSet Fairness Subscore**: StereoSet fornisce un "Overall Score" che bilancia accuracy e stereotypicality. Questo può essere integrato come una misura di fairness bilanciata con accuratezza.
* **WinoBias Accuracy Subscore**: Valuta se il modello risolve correttamente i pronomi, tenendo conto della possibilità di bias di genere.
* **CrowS-Pairs Neutral Preference**: Basato sulla percentuale di volte che il modello preferisce la frase neutra o non stereotipata, indicando la sua capacità di evitare il bias mantenendo la precisione linguistica.

****

Questa formula combina la correttezza del modello con la sua capacità di mantenere un comportamento equo e neutro.

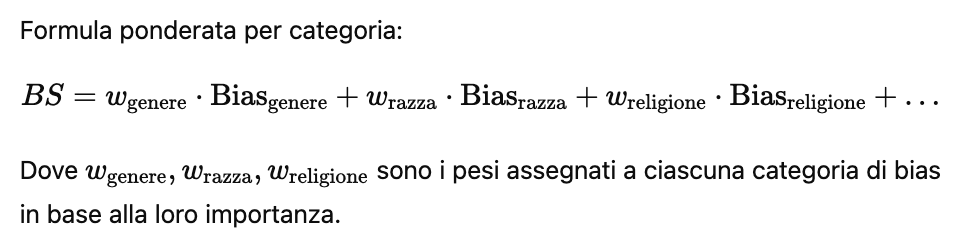
### 3. Metrica Finale Unificata: ****Unified Bias-Accuracy Score (UBAS)****

Il punteggio finale del framework sarà una combinazione del **Bias Score** (BS) e del **Fairness-Accuracy Score** (FA). Questo bilancerà la capacità del modello di evitare bias con la sua accuratezza generale.

****

### 4. Ponderazione per Categorie di Bias

Un'altra considerazione importante è la possibilità di **ponderare i bias** in base alle categorie sociali. Ad esempio, se si ritiene che il bias di genere e di razza siano più critici, si potrebbero assegnare pesi diversi ai punteggi relativi a queste categorie.

****

### 5. Valutazione Globale

Alla fine, **UBAS** fornirà un punteggio che bilancia sia l'accuratezza linguistica che l'equità del modello, offrendo un modo robusto per confrontare diversi modelli su un ampio spettro di bias e task linguistici.

### 6. Implementazione

Per implementare questo framework, dovresti:

* **Normalizzare i punteggi** di ciascun benchmark per renderli comparabili, poiché le scale di valutazione potrebbero essere diverse.
* Definire chiaramente i pesi da assegnare a ciascuna categoria di bias o task, in base all'importanza che desideri attribuire a ciascun aspetto.
* Implementare una pipeline automatizzata che esegue i test su ciascun modello, raccoglie i punteggi dai diversi benchmark, li normalizza e calcola il punteggio unificato.

### Conclusione

Unire **StereoSet**, **CrowS-Pairs** e **WinoBias** sotto un'unica metrica è fattibile e potrebbe portare a una valutazione più completa e robusta del bias nei modelli di intelligenza artificiale. La metrica unificata **UBAS** bilancia il livello di bias del modello con la sua accuratezza, permettendo una valutazione equa delle sue performance. La chiave del successo di questo approccio è una corretta normalizzazione dei punteggi e una ponderazione bilanciata dei bias e dei task linguistici.

**Step 2: Rappresentazione Multivettoriale delle Frasi Generate dalle LLM**

**Descrizione**: Applica modelli di linguaggio di grandi dimensioni (LLM) come GPT-4 per generare rappresentazioni multivettoriali delle frasi. Integra i KGs per creare vettori ibridi, combinando embedding di frasi e relazioni tra concetti.

**Come utilizzare i KGs**: Durante la fase di rappresentazione multivettoriale, arricchisci le frasi con **embedding derivati dai KG**. In altre parole, non limitarti agli embedding delle parole (come quelli di GPT-4), ma aggiungi anche rappresentazioni che tengano conto delle **relazioni tra le entità** presenti nei KGs.

**Esempio**:

* Per una frase come "Angela Merkel è stata una grande leader", oltre all'embedding della frase generato da GPT-4, puoi utilizzare un KG come Wikidata per aggiungere un embedding relativo alla persona "Angela Merkel", con informazioni come **ruolo (ex-cancelliera)**, **nazionalità (tedesca)**, **periodo storico** e **politica di genere**. Questi dati aiutano a creare una rappresentazione più ricca della frase.

### 1. ****Utilizzo degli Embedding Multivettoriali per l'Analisi del Bias****

Gli embedding multivettoriali permettono di integrare le informazioni testuali (cioè la formulazione della frase) con le informazioni semantiche (cioè le relazioni tra le entità nella frase). Questo è particolarmente utile per rilevare **bias complessi** che potrebbero non essere evidenti solo dal testo, ma che emergono quando consideri anche le relazioni tra concetti (fornite dai KGs).

#### a. **Identificare i Bias con gli Embedding**

Gli embedding multivettoriali contengono informazioni che si possono analizzare per scoprire bias. Ad esempio:

* Bias di genere nei ruoli professionali: L'integrazione delle informazioni testuali e semantiche ti permette di vedere come ruoli come "ingegnere" o "dottore" sono associati a generi specifici (in base alle relazioni estratte dai KGs).
* Bias etnico o culturale: Gli embedding semantici possono rivelare bias impliciti in relazioni storiche tra gruppi etnici, eventi storici o stereotipi culturali.

**Esempio**: Per una frase come "Le donne sono brave a cucinare":

* **Embedding testuale**: catturerà il significato linguistico della frase, che evidenzia uno stereotipo.
* **Embedding semantico**: potrebbe aggiungere informazioni su come il ruolo della cucina è storicamente associato alle donne in certi contesti culturali.

Si possono usare gli embedding combinati per rilevare se una frase è problematicamente stereotipata o meno.

#### b. **Algoritmi di Analisi del Bias**

Si possono usare gli embedding come input per **algoritmi di analisi del bias**, che rilevano se una frase è associata a determinati bias. Ad esempio, un classificatore addestrato su dati etichettati con bias può prevedere se una nuova frase contiene bias di genere, razza o cultura.

### 2. ****Sviluppo di Modelli di Classificazione****

Gli **embedding multivettoriali** possono essere usati come input per sviluppare modelli di **classificazione**, che ti permettono di:

* **Classificare frasi** in base alla presenza o assenza di bias.
* **Prevedere categorie** di bias (ad esempio, bias di genere, bias razziale, bias professionale).

#### a. **Rete Neurale per la Classificazione del Bias**

Uno degli approcci più comuni è usare una **rete neurale** per classificare le frasi. Puoi prendere l'embedding multivettoriale di una frase e passarlo attraverso una rete neurale per determinare se quella frase contiene bias.

L'output sarà una previsione binaria:

* **Bias** (1)
* **Non-bias** (0)

#### b. **Addestramento del Modello**

Addestrare il modello di classificazione con un dataset etichettato (come StereoSet), in cui ogni frase è associata a una classe che indica se contiene o meno bias.

#### c. **Valutazione del Modello**

Si possono misurare le prestazioni del modello con metriche di classificazione standard come **accuratezza**, **precisione**, **richiamo** e **F1-score**.

### 3. ****Clusterizzazione delle Frasi con Embedding Multivettoriali****

La **clusterizzazione** è un'altra tecnica utile per analizzare gli embedding multivettoriali. Con la clusterizzazione, si possono raggruppare frasi che presentano **bias simili** in termini sia testuali che semantici. Questo permette di identificare categorie di bias che potrebbero non essere immediatamente evidenti.

#### a. **Algoritmi di Clusterizzazione**

Si possono utilizzare algoritmi di clusterizzazione come **K-Means** o **Agglomerative Clustering** per raggruppare le frasi sulla base della loro rappresentazione multivettoriale.

#### b. **Interpretazione dei Cluster**

I cluster risultanti ti permettono di vedere gruppi di frasi che condividono similarità nel modo in cui il bias si manifesta. Ad esempio, si potrebbe trovare un cluster di frasi che trattano stereotipi di genere nelle professioni, o un cluster che evidenzia bias razziali legati a determinati contesti geografici.

### 4. ****Validazione e Interpretabilità del Modello****

È importante validare il comportamento del modello e capire **come** e **perché** il modello ha preso determinate decisioni. Per farlo, si possono utilizzare tecniche di **interpretabilità** come:

#### a. **SHAP (SHapley Additive exPlanations)**

SHAP è una tecnica che permette di capire l'impatto di ogni feature (in questo caso, dimensioni degli embedding) sulla decisione del modello.

Questo permette di vedere **quali aspetti degli embedding** (testuali o semantici) hanno influenzato di più il modello nel classificare una frase come biasata o non biasata.

#### b. **LIME (Local Interpretable Model-agnostic Explanations)**

LIME è un'altra tecnica che si può usare per spiegare le decisioni del modello. Genera spiegazioni locali, mostrando quali parti dell'embedding hanno influenzato la decisione per ciascun esempio.

### Come Implementare lo Step 2 con StereoSet

1. **Estrarre Entità da StereoSet**:
   * Usare strumenti come **Named Entity Recognition (NER)** per identificare entità chiave nelle frasi di StereoSet (ad esempio, persone, professioni, gruppi etnici, nazionalità).
2. **Arricchire le Entità con KGs**:
   * Una volta identificate le entità, usa Knowledge Graphs come **Wikidata** o **DBpedia** per estrarre informazioni aggiuntive e relazioni tra le entità.
   * **Esempio**: Per una frase che menziona una professione, usare il KG per trovare dati storici sulla predominanza di genere in quella professione.
3. **Creare Embedding Multivettoriali**:
   * Generare **embedding testuali** usando modelli pre-addestrati come **BERT** o **GPT-4**.
   * Integrare le informazioni semantiche ottenute dai KGs creando **embedding semantici**.
   * Combinare le due rappresentazioni in un **vettore ibrido** che rappresenta la frase.

### Passaggi per Creare Embedding Multivettoriali:

### 1. ****Generare Embedding Testuali con LLM****

Il primo passo consiste nel generare embedding testuali utilizzando un modello di linguaggio pre-addestrato come **BERT** o **GPT-4**.

#### a. **Input: Le frasi di StereoSet**

StereoSet contiene frasi stereotipiche e frasi neutre. Per ogni frase di StereoSet, generare una rappresentazione numerica (embedding) che catturi il significato della frase.

#### b. **Modelli di Linguaggio (LLM)**

Utilizzare modelli come **BERT** o **GPT-4** per ottenere embedding testuali. Questi modelli sono addestrati su grandi quantità di testo e sono in grado di creare vettori che rappresentano la semantica delle frasi.

* **BERT** (Bidirectional Encoder Representations from Transformers): Per ottenere embedding con BERT, si può usare la versione pre-addestrata del modello disponibile in librerie come **Hugging Face Transformers**. BERT genera rappresentazioni bidirezionali, che tengono conto del contesto delle parole sia prima che dopo di esse nella frase.
* **GPT-4** (Generative Pre-trained Transformer): GPT-4 è un modello più recente, basato su architettura transformer, che genera rappresentazioni molto ricche delle frasi. Anche questo modello può essere utilizzato per creare embedding che rappresentano il contesto e il significato della frase.

#### c. **Come Generare gli Embedding**

Utilizzando una delle librerie LLM, per esempio **Hugging Face**, si può ottenere un embedding per ciascuna frase di StereoSet.

**Output**: L'output sarà un vettore numerico (embedding testuale) che rappresenta la frase.

### 2. ****Creare Embedding Semantici usando i Knowledge Graphs (KGs)****

Il secondo passo è arricchire queste frasi con informazioni semantiche dai **Knowledge Graphs (KGs)**, come **Wikidata** o **DBpedia**, che forniscono relazioni tra entità menzionate nelle frasi.

#### a. **Identificare le Entità nella Frase**

Per creare embedding semantici, prima **estrarre le entità** rilevanti dalle frasi di StereoSet. Questo può essere fatto tramite tecniche di **Named Entity Recognition (NER)**, che identificano entità come persone, luoghi, organizzazioni, professioni, ecc.

**Output**: Le entità identificate potrebbero essere:

* "Angela Merkel" (PERSONA)
* "Germania" (LUOGO)

#### b. **Ottenere le Relazioni dai KGs**

Una volta identificate le entità, puoi usare un KG come **Wikidata** o **DBpedia** per ottenere informazioni strutturate e relazioni tra queste entità. Lo si può fare usando query **SPARQL** per interrogare i KGs e ottenere le informazioni desiderate.

Questa query restituisce le proprietà e le relazioni di "Angela Merkel" nel contesto di **Wikidata**, come:

* "Cancelliere della Germania"
* "Politico"
* Periodo storico associato, ecc.

#### c. **Creare Embedding Semantici dalle Relazioni**

Usare le relazioni estratte dai KGs per creare un **vettore semantico** che rappresenta le connessioni tra le entità nella frase.

Esempio: Per la frase "Angela Merkel è stata una grande leader in Germania", l'**embedding semantico** potrebbe includere vettori che rappresentano:

* La relazione tra **Angela Merkel** e la **carica politica** (cancelliere della Germania).
* Il ruolo di Angela Merkel nel contesto di **eventi storici** (relazioni temporali).

Si può rappresentare queste relazioni come vettori numerici o come feature aggiuntive da combinare con l'embedding testuale.

### 3. ****Combinare gli Embedding Testuali e Semantici in un Vettore Ibrido****

L'ultimo passo è **combinare** gli embedding testuali ottenuti dai LLM con quelli semantici derivati dai KGs. Questo processo di combinazione permette di ottenere un **embedding multivettoriale** che contiene sia il significato linguistico della frase che le sue relazioni concettuali o storiche.

#### a. **Concatenazione degli Embedding**

Un modo semplice per combinare questi due tipi di embedding è la **concatenazione** dei vettori. Se hai un embedding testuale di dimensione 768 (ad esempio da BERT) e un embedding semantico di dimensione 300 (da relazioni nel KG), puoi concatenarli in un unico vettore di dimensione 1068.

#### b. **Ponderazione degli Embedding (opzionale)**

In alcuni casi, si potrebbe voler dare più peso a una delle due rappresentazioni (testuale o semantica), a seconda dell'importanza. Ad esempio, se le relazioni semantiche del KG sono particolarmente rilevanti per l'analisi del bias, puoi dare loro un peso maggiore.

**Step 3: Clusterizzazione Gerarchica con Validazione**

**Descrizione**: Applica una clusterizzazione gerarchica sui risultati precedenti e utilizza tecniche di validazione come **Silhouette Score** o **Davies-Bouldin Index** per valutare la qualità dei cluster. Usa KGs per arricchire i cluster e validare i risultati in base a relazioni semantiche.

**Come utilizzare i KGs**: I KGs possono fornire ulteriori informazioni che aiutano a validare la coerenza dei cluster in base alle relazioni tra entità e concetti.

**Esempio**:

* Se un cluster contiene frasi legate a stereotipi di genere, il KG può aiutare a confermare la validità del cluster mostrando che le entità nel cluster sono tutte collegate a contesti di bias storici o culturali. Questo può servire come ulteriore validazione semantica dei cluster.

### Tale tecnica è utile per: 1. ****Scoperta e Identificazione di Bias Latenti****

Il **clustering gerarchico** permette di esplorare i dati testuali o gli embedding multivettoriali a **diversi livelli di granularità**, rivelando potenziali bias che potrebbero non essere evidenti a prima vista. Usando la validazione, possiamo assicurarci che i cluster formati siano **coerenti** e **significativi**.

#### a. **Bias Latenti nei Cluster**

Il clustering gerarchico aiuta a scoprire **bias latenti** nei modelli di linguaggio:

* Cluster più grandi possono rivelare **bias generali** (es. bias di genere in ruoli professionali).
* Sottocluster possono rivelare **bias specifici** (es. bias legati a **specifiche professioni** come ingegneri o medici).

**Esempio**: In un cluster di frasi riguardanti il lavoro, potresti trovare che le frasi che parlano di donne sono maggiormente associate a ruoli di supporto, mentre quelle sugli uomini sono associate a ruoli di leadership.

#### b. **Validazione della Coerenza dei Cluster**

La **validazione** del clustering permette di verificare se i cluster risultanti sono coerenti e ben definiti. Utilizzando metriche come il **Silhouette Score** o il **Davies-Bouldin Index**, puoi valutare se i cluster trovati hanno confini chiari e se i dati sono ben separati. Se i cluster che rivelano bias sono ben distinti, sei più sicuro dell’ analisi.

### 2. ****Scalabilità dell'Analisi del Bias****

Il clustering gerarchico è utile per **scalare l'analisi del bias** in dataset di grandi dimensioni. Invece di analizzare **singole frasi** una per una, utilizzare il clustering gerarchico per raggruppare frasi o concetti simili in **cluster tematici**, e poi validare la coerenza di quei gruppi.

#### a. **Raggruppare Frasi per Argomento o Bias**

Applicare il clustering gerarchico per identificare cluster che contengono frasi con **bias simili**. Una volta creati i cluster, utilizzare strumenti di validazione per garantire che i gruppi di frasi siano **logicamente coerenti**.

**Esempio**: Se un cluster contiene frasi stereotipate su **ruoli di genere** o **etnia**, puoi validare la consistenza di quel cluster e approfondire l'analisi per vedere se questi bias sono presenti in modo significativo in quel gruppo.

### 3. ****Flessibilità e Scoperta di Pattern Complessi****

Il clustering gerarchico offre la possibilità di **esplorare i dati a diversi livelli** di granularità. Eseguire un'analisi flessibile, dove scopri pattern di bias **più generali** o **più specifici**, e con la **validazione** puoi determinare il livello di profondità ottimale per l'analisi.

#### a. **Scoperta di Bias Multi-livello**

Poiché il clustering gerarchico crea una struttura ad albero (dendrogramma), identificare che il bias sia a livello **superficiale** (grandi cluster) che a livello **più profondo** (sottocluster). Questo è utile per esplorare **bias più sottili** o **combinazioni di bias** che emergono solo quando si considerano relazioni più complesse.

**Esempio**:

* A un livello superficiale, puoi identificare bias di genere nelle professioni.
* A un livello più profondo, potresti scoprire bias più sottili, come la rappresentazione delle donne in specifici **ruoli professionali** (es. "insegnante" o "medico").

#### b. **Validazione per Identificare il Numero Ottimale di Cluster**

Dopo aver eseguito il clustering gerarchico, utilizzare tecniche di **validazione** per determinare il numero ottimale di cluster. Questo è importante perché il **taglio del dendrogramma** influisce sul numero finale di cluster che utilizzi per l'analisi.

* **Silhouette Score**: Valutare quanto ogni punto in un cluster è simile ai punti nel proprio cluster rispetto ai punti negli altri cluster.
* **Davies-Bouldin Index**: Misurare la separabilità tra cluster; più basso è il punteggio, migliore è la separazione dei cluster.

La validazione consente di tagliare il dendrogramma al livello ottimale, ottenendo così il miglior bilanciamento tra numero di cluster e coerenza interna.

### 4. ****Identificazione e Mitigazione del Bias Sistemico****

Utilizzare il clustering gerarchico con validazione permette anche di scoprire **bias sistemici**, ovvero pattern di bias che si ripetono su larga scala all'interno del modello di linguaggio o dei dati. La validazione aiuta a verificare se questi pattern sono **consistenti** e **significativi** o se sono semplicemente anomalie.

#### a. **Bias Ripetuti in Cluster Diversi**

Con il clustering gerarchico vedere se certi **bias sistemici** (come stereotipi di genere o razziali) si ripetono attraverso cluster distinti. La validazione garantisce che questi cluster siano ben definiti e che il bias individuato sia **robusto**.

**Esempio**:

* Se diversi cluster che riguardano professioni mostrano un **bias di genere** coerente (es. frasi che associano uomini a ruoli tecnici e donne a ruoli di supporto), verificare se il bias è sistematico e non casuale.

#### b. **Mitigazione del Bias**

Dopo aver identificato questi bias sistemici, applicare tecniche di **mitigazione del bias**, come la **regolarizzazione** o il **debiasing degli embedding**. La validazione del clustering aiuta a capire **dove** intervenire e quanto il bias è diffuso nei dati.

### 5. ****Monitoraggio e Valutazione delle Tecniche di Mitigazione del Bias****

Una volta applicate tecniche di **mitigazione del bias** (come il debiasing degli embedding o il disentanglement), riapplicare il clustering gerarchico con validazione per verificare se il bias è stato **ridotto** o **eliminato**. La validazione ti permette di misurare l'efficacia delle tecniche di mitigazione.

#### a. **Rieseguire il Clustering dopo la Mitigazione**

Dopo aver applicato tecniche di mitigazione, eseguire nuovamente il clustering gerarchico e usare strumenti di validazione per verificare se i cluster che in precedenza mostravano bias evidenti sono ora **più neutri** o **equilibrati**.

**Esempio**:

* Se un cluster che in precedenza mostrava bias di genere ora ha una **distribuzione più equilibrata** di frasi e sentimenti, questo indica che la mitigazione ha avuto successo.

**Step 4: Analisi Sentimentale sui Cluster e sulle frasi/vettori**

**Descrizione**: Applicare strumenti di analisi del sentiment come **VADER** o **TextBlob** sui cluster creati, tenendo in considerazione i risultati dell’incrocio con altre fonti di dati.

Utilizzare i KGs per arricchire l'analisi del sentiment, incorporando le relazioni semantiche tra le entità.

**Come utilizzare i KGs**: I KGs possono aggiungere informazioni contestuali per aiutare a determinare il sentiment di frasi ambigue, basandosi su relazioni esplicite tra entità e concetti.

**Esempio**:

* In un cluster che contiene frasi come "Gli immigrati sono una minaccia", l'analisi del sentiment potrebbe rilevare una connotazione negativa. Usando un KG, puoi rafforzare l'analisi del sentiment identificando relazioni tra il termine "immigrati" e concetti come "pregiudizi" o "discriminazione". Questo migliora la capacità del sistema di rilevare bias impliciti negativi.

N.B. è possibile adottare altre tecniche più potenti come ad esempio WEAT o VADER. Di seguito ne elenco alcune:

### Tecniche Aggiuntive di Rilevamento del Bias

#### 1. **Word Embedding Association Test (WEAT)**

Il **Word Embedding Association Test (WEAT)** è una tecnica che viene utilizzata per misurare il **bias semantico** negli **embedding**. Questo test confronta la **similarità** tra gruppi di parole associate a categorie sensibili (come "uomo", "donna", "carriera", "famiglia") e misura se certi gruppi sono più strettamente associati a concetti stereotipati.

##### a. **Applicazione sugli Embedding**

* **Obiettivo**: Misurare se gli embedding delle parole o frasi contenenti entità sensibili mostrano un bias rispetto a categorie specifiche.
* **Come funziona**: Calcolare la distanza tra embedding di parole sensibili (es. "uomo", "donna") e concetti stereotipati (es. "carriera", "famiglia"). Se gli embedding associati a "uomo" sono più vicini a concetti legati al lavoro e quelli di "donna" sono più vicini a concetti legati alla famiglia, questo indica un bias.

**Vantaggi**:

* Rilevare bias impliciti e **misura l'effetto del bias** direttamente sugli embedding.
* Fornisce un'**analisi quantitativa** del bias, basata sulla similarità tra parole.

#### 2. **Counterfactual Fairness**

La **Counterfactual Fairness** (o equità controfattuale) è una tecnica che permette di verificare se cambiando un attributo sensibile (es. genere, etnia) in una frase, il modello produce una **previsione diversa**. Questo aiuta a rilevare se il bias è causato dall'uso di caratteristiche sensibili.

##### a. **Applicazione sulle Frasi**

* **Obiettivo**: Rilevare bias testuale generando frasi controfattuali in cui l'attributo sensibile viene cambiato e verificando se il significato o il sentimento del testo cambia.
* **Come funziona**: Cambiare attributi sensibili nelle frasi (es. sostituisci "uomo" con "donna") e confronta il risultato. Se la modifica cambia significativamente il significato o il sentiment della frase, potrebbe esserci un bias.

**Esempio**:

1. **Frase originale**: "Il CEO ha preso una decisione forte."
2. **Frase controfattuale**: "La CEO ha preso una decisione forte."

Confrontando il **sentiment** o la **predizione** su queste due frasi, è possibile rilevare se il modello risponde in modo diverso in base al genere del soggetto.

**Vantaggi**:

* Permette di **testare direttamente l'impatto di attributi sensibili** nel comportamento del modello.
* È particolarmente utile per rilevare **bias di genere**, **etnia**, o altri attributi personali.

#### 3. **Fairness Through Awareness**

Questa tecnica si basa sull'assicurarsi che il modello sia consapevole delle **differenze contestuali** tra gruppi diversi, ma senza utilizzare queste informazioni per prendere decisioni discriminatorie. Consiste nell'**includere conoscenze contestuali** nei modelli per evitare conclusioni stereotipate.

##### a. **Applicazione sugli Embedding e Frasi**

* **Obiettivo**: Aggiungere consapevolezza delle differenze culturali, etniche o di genere nei modelli di linguaggio senza che queste informazioni influenzino le previsioni in modo discriminatorio.
* **Come funziona**: Usare i **Knowledge Graphs** per arricchire gli embedding multivettoriali con informazioni contestuali (es. storia culturale di una professione), ma separa queste informazioni durante le previsioni per evitare che influenzino decisioni legate a bias.

**Esempio**: Arricchire gli embedding di professioni come "medico" o "insegnante" con informazioni aggiuntive tratte da Knowledge Graphs (ad es., la predominanza di genere in quelle professioni in contesti culturali diversi), ma si utilizzano tecniche di **disentanglement** per evitare che il modello colleghi queste informazioni a predizioni di bias.

**Vantaggi**:

* Consente di aggiungere **contesto storico o culturale** alle frasi e agli embedding senza introdurre bias discriminatorio.
* Riduce il rischio di decisioni errate basate su stereotipi impliciti.

#### 4. **BiasFinder: Test Metamorfico**

Il **BiasFinder** è una tecnica che genera **test metamorfico** per individuare bias nei modelli di linguaggio. I test metamorfici sono trasformazioni di input che dovrebbero produrre risultati simili se il modello è equo. Se la risposta del modello cambia in modo significativo, ciò potrebbe indicare un bias.

##### a. **Applicazione su Embedding e Cluster**

* **Obiettivo**: Generare frasi alternative (o metamorfosi) che dovrebbero essere trattate allo stesso modo dal modello, e confrontare il risultato.
* **Come funziona**: Generare versioni leggermente modificate delle frasi in un cluster e verifica se il modello fornisce risposte diverse o produce embedding significativamente differenti. Se il modello risponde in modo diverso a frasi che dovrebbero essere trattate allo stesso modo, potrebbe esserci un bias.

**Esempio**:

1. **Frase originale**: "Il dottore ha curato il paziente."
2. **Frase metamorfica**: "Il medico ha curato il paziente."

Se il modello produce un risultato diverso per queste due frasi, potrebbe indicare un bias basato sull'uso di terminologie specifiche.

**Vantaggi**:

* Permette di **testare la robustezza** del modello contro piccole variazioni nella terminologia o nella struttura delle frasi.
* Rilevare bias non solo nelle predizioni, ma anche negli embedding generati dal modello.

#### 5. **Analisi della Distribuzione degli Attributi Sensibili nei Cluster**

Un'altra tecnica utile per rilevare il bias nei **cluster** è l'**analisi della distribuzione degli attributi sensibili**. Dopo aver creato i cluster di frasi, è possibile analizzare se certi **attributi sensibili** (es. genere, etnia) sono distribuiti in modo non equilibrato all'interno dei cluster.

##### a. **Applicazione sui Cluster**

* **Obiettivo**: Verificare se certi cluster contengono una sovrabbondanza di frasi con attribuzioni a certi gruppi sensibili (es. solo frasi su donne in ruoli domestici).
* **Come funziona**: Contare le occorrenze di entità legate a attributi sensibili all'interno dei cluster e verifica se certi gruppi sono sovra o sotto-rappresentati in alcuni cluster.

**Esempio**: Se un cluster contiene prevalentemente frasi su **donne** in ruoli domestici e un altro contiene prevalentemente frasi su **uomini** in ruoli professionali, questo potrebbe rivelare uno squilibrio o un bias stereotipico.

**Vantaggi**:

* Ti permette di rilevare **bias sistemici** nella rappresentazione di certi gruppi nei cluster.
* Fornisce una visione chiara di come il modello sta raggruppando entità sensibili.

### Riassunto: Dove Applicare le Tecniche di Rilevamento del Bias

* **Sulle frasi**: Applicare tecniche come la **Counterfactual Fairness** e il **BiasFinder** per testare come il modello tratta frasi con attributi sensibili modificati.
* **Sugli embedding multivettoriali**: Utilizzare il **WEAT** per rilevare bias semantici negli embedding e il **Fairness Through Awareness** per assicurarti che il contesto non introduca bias.
* **Sui cluster**: Analizzare la **distribuzione degli attributi sensibili** nei cluster e applicare l'**analisi sentimentale aggregata** per capire se gruppi di frasi mostrano pattern di bias emotivo.

**MODULO MITIGATORIO**

**Step 5: Integrazione tra Clusterizzazione e Rete Avversaria**

**Descrizione**: Integra reti avversarie per identificare e correggere bias durante il processo di clusterizzazione. I KGs possono aiutare la rete avversaria a rilevare e correggere bias in tempo reale.

**Come utilizzare i KGs**: La rete avversaria può utilizzare le informazioni dei KGs per migliorare il rilevamento dei bias durante la clusterizzazione, riconoscendo **pattern storici** o **relazioni culturali** problematiche.

**Esempio**:

* Se la rete avversaria rileva che un cluster contiene frasi che discriminano un gruppo etnico, il KG può fornire contesto sulle relazioni storiche tra i gruppi e suggerire come correggere il bias nei nuovi output del modello, in modo dinamico e adattivo.

**Aggiunta di Tecniche di Disentanglement + Regularization**

**Descrizione**: Integrare tecniche di disentanglement per separare le rappresentazioni latenti degli attributi sensibili (genere, etnia, ecc.) dalle altre caratteristiche semantiche. Usare KGs per migliorare questo processo, separando le informazioni neutrali dalle caratteristiche che potrebbero influenzare il bias.

**Come utilizzare i KGs**: I KGs permettono di ottenere una rappresentazione chiara delle **categorie sensibili** e delle **relazioni tra le entità**. Queste possono essere utilizzate per isolare in modo più accurato gli attributi sensibili dai restanti vettori.

**Esempio**:

* Se una frase contiene il termine "casalinga", il KG può aiutare a identificare che questo termine è storicamente legato al genere femminile e potrebbe essere associato a un bias. Il modello può quindi essere regolarizzato per non collegare il concetto di "lavoro domestico" esclusivamente al genere femminile.

**Obiettivo del Disentanglement**

L'obiettivo principale del **disentanglement** è di assicurare che le rappresentazioni latenti create dal modello (ad esempio, gli **embedding**) **non riflettano indebitamente attributi sensibili**, come il genere o l'etnia, quando questi attributi non dovrebbero influenzare la predizione o il comportamento del modello. Separando chiaramente gli attributi sensibili dalle altre caratteristiche latenti, possiamo:

1. **Ridurre il bias** nelle previsioni.
2. **Migliorare l'equità** del modello, assicurando che esso non prenda decisioni basate su attributi che non dovrebbero influenzare il risultato.

**Esempi di Bias e Disentanglement**

1. **Bias di Genere nei Ruoli Professionali**:
   * Se un modello di linguaggio associa il **genere** (es. "uomo" o "donna") a certi ruoli professionali in modo stereotipato, come "ingegnere" per uomini e "infermiera" per donne, significa che il modello sta **mescolando** l'informazione del **genere** con le caratteristiche semantiche della professione.
   * **Disentanglement** permette di separare la rappresentazione del **genere** da quella della **professione** nel modello, in modo che la professione possa essere rappresentata senza che il genere influisca in modo inappropriato sulla previsione.
2. **Bias di Etnia nella Previsione di Emozioni**:
   * Se il modello predice il **sentimento** o l'**emozione** di una frase e usa **etnia** o **origine culturale** per influenzare questa predizione, potrebbe amplificare stereotipi negativi (es. associando emozioni di **rabbia** o **criminalità** a determinate etnie).
   * Il **disentanglement** separa l'**etnia** dal **contesto emotivo** della frase, impedendo al modello di associare implicitamente l'etnia a connotazioni emotive inappropriate.

**Vantaggi e Utilità del Disentanglement**

**1. Riduzione del Bias**

Separare le rappresentazioni latenti aiuta a evitare che il modello introduca **bias impliciti** durante le predizioni. Ad esempio, se un attributo come il **genere** o l'**etnia** è mescolato nella rappresentazione con altre caratteristiche, il modello può utilizzare queste informazioni in modo non corretto, portando a previsioni che riflettono bias culturali o sociali.

**Esempio**: In un sistema di raccomandazione di lavoro, se gli embedding latenti per "donna" sono collegati troppo strettamente con professioni storicamente dominate da donne (come l'insegnamento o il lavoro domestico), il modello potrebbe **limitare le raccomandazioni** per donne a tali ruoli. Il **disentanglement** permette di slegare il genere dal tipo di raccomandazione, rendendo il sistema più **equo**.

**2. Interpretabilità e Trasparenza**

Le tecniche di disentanglement migliorano anche l'**interpretabilità** del modello. Separando gli attributi sensibili da quelli neutri, è più facile capire come il modello sta prendendo le decisioni e verificare se stia usando in modo inappropriato attributi sensibili per fare previsioni.

**Esempio**:

* Un modello di classificazione di testo che predice se un commento è **positivo** o **negativo** può essere interpretato meglio se sappiamo che la rappresentazione latente della **razza** o del **genere** non è mescolata con il contenuto emotivo del commento. In questo modo possiamo essere sicuri che il modello non stia associando sentimenti negativi a certe razze o generi.

**3. Equità e Non Discriminazione**

Separare le rappresentazioni latenti degli attributi sensibili migliora la **fairness** (equità) dei modelli. In scenari come le assunzioni o le raccomandazioni di prodotti, è fondamentale che attributi come il genere, l'etnia o l'età **non influenzino** il risultato. Il disentanglement permette al modello di considerare questi attributi separatamente e di **non discriminare** basandosi su di essi.

**Esempio**: In un sistema di reclutamento, se gli embedding di un candidato contengono informazioni latenti sul suo **genere** o **etnia**, il sistema potrebbe discriminare in modo inconsapevole, suggerendo posizioni solo per certi gruppi. Separando tali attributi, il sistema si concentra sulle **competenze** piuttosto che sugli attributi sensibili.

**4. Debiasing del Modello**

Le tecniche di disentanglement possono essere utilizzate anche per applicare tecniche di **debiasing**. Dopo aver separato le rappresentazioni, si può agire direttamente su come il modello utilizza o non utilizza gli attributi sensibili, limitando l'influenza di questi ultimi nelle decisioni.

**Esempio**: In un modello che valuta la **probabilità di approvazione di un prestito**, è essenziale evitare che attributi come **etnia** o **età** influiscano direttamente nella decisione. Attraverso il disentanglement, puoi slegare tali attributi dalla rappresentazione finanziaria, garantendo che non siano utilizzati in modo inappropriato per determinare l'approvazione.

**Tecniche Usate per il Disentanglement**

Per implementare il **disentanglement** delle rappresentazioni latenti, esistono diverse tecniche che separano i fattori sensibili dai fattori latenti:

**1. Variational Autoencoders (VAE) con Disentanglement**

Un approccio comune è l'uso di **Variational Autoencoders (VAE)** per imparare una rappresentazione latente in cui gli attributi sensibili sono separati dalle altre variabili. Alcuni modelli di VAE includono componenti che forzano la separazione delle informazioni latenti, minimizzando la correlazione tra attributi sensibili e latenti.

* **Come funziona**: Un VAE prende input (frasi, embedding) e genera uno spazio latente che contiene tutte le informazioni utili. Con **disentanglement**, si applicano restrizioni in modo che il modello impari a tenere separati gli attributi sensibili, come genere o etnia, da altre caratteristiche semantiche.

**2. Adversarial Debiasing (Reti Avversarie)**

Le **reti avversarie** possono essere usate per eliminare gli attributi sensibili dagli embedding latenti. Un **discriminatore** tenta di identificare gli attributi sensibili dalla rappresentazione latente, mentre il modello tenta di imparare una rappresentazione che **nasconda** tali attributi al discriminatore.

* **Come funziona**: Durante l'addestramento, una rete avversaria cerca di identificare attributi sensibili (genere, etnia) nelle rappresentazioni latenti. Il modello è penalizzato se il discriminatore riesce a identificare tali attributi, incentivando così il modello a produrre embedding che non contengano informazioni sugli attributi sensibili.

**3. Feature Regularization**

L'applicazione di **penalità regolarizzatrici** può aiutare a limitare l'influenza degli attributi sensibili. Attraverso tecniche di **feature regularization**, si può penalizzare l'uso di informazioni legate a variabili come il genere o l'etnia, forzando il modello a fare predizioni basate solo su attributi **neutri** o **pertinenti**.

* **Come funziona**: Aggiungi un termine di regolarizzazione alla funzione di perdita del modello, che penalizza il modello se le rappresentazioni latenti dipendono fortemente dagli attributi sensibili. Questo costringe il modello a ridurre l'uso di attributi come genere o etnia nelle sue decisioni.

**Conclusione: L'Utilità del Disentanglement**

In sintesi, l'integrazione di tecniche di **disentanglement** è estremamente utile per:

1. **Ridurre il bias** nei modelli di linguaggio, assicurandoti che attributi sensibili come **genere** o **etnia** non influenzino le decisioni.
2. Migliorare la **trasparenza** e l'**interpretabilità** del modello, separando attributi sensibili da altre variabili.
3. Garantire **equità** nei processi decisionali, riducendo la possibilità di discriminazione.
4. Applicare tecniche di **debiasing** in modo mirato, eliminando l'influenza degli attributi sensibili sulle predizioni.

#### **Come si collega lo Step 5 allo Step 4**

1. **Input dalla fase precedente (Step 4):**
   * **Cluster e frasi con bias o sentiment negativi**: Dallo step 4, ricevi un elenco di frasi o cluster che hanno mostrato bias espliciti o impliciti. Questi sono i target della rete avversaria, che deve imparare a rilevarli e correggerli.
   * **KGs arricchiti**: Le informazioni aggiuntive sui concetti e sulle relazioni semantiche tra entità derivano dai KGs. Queste informazioni vengono passate alla rete avversaria per migliorare la sua capacità di riconoscere bias.
2. **Processo della rete avversaria (Step 5):**
   * **Struttura della rete avversaria**: La rete avversaria, o GAN (Generative Adversarial Network), ha due componenti principali:
     + **Generatore**: Cerca di generare frasi o cluster privi di bias partendo dai dati di input. Questo potrebbe avvenire creando rappresentazioni latenti delle frasi (vettori) in cui i bias rilevati nello step 4 vengono ridotti.
     + **Discriminatore**: Il discriminatore cerca di **rilevare il bias** nelle frasi o nei cluster generati dal generatore, confrontandoli con i dati di input arricchiti di bias dallo step 4. Il discriminatore, grazie alle informazioni fornite dai KGs, può diventare più sensibile ai bias di contesto culturale, storico o sociale.
   * **Rete Adversaria con KGs**: La rete usa i **KGs** come fonte di informazione aggiuntiva per il discriminatore. Ad esempio, se un cluster contiene frasi con pregiudizi storici o culturali, il KG può informare il discriminatore del legame tra certe entità e concetti problematici, migliorando la capacità della rete di riconoscere bias più sottili.
   * **Disentanglement dei bias**: Durante l'addestramento, la rete avversaria può usare tecniche di **disentanglement** per separare gli attributi sensibili (come genere, etnia, ecc.) dalle altre caratteristiche semantiche delle frasi, in modo da isolare i bias e correggerli senza alterare il significato complessivo.
3. **Output della rete avversaria (Step 5):**
   * **Frasi o cluster privi di bias**: Il generatore, attraverso il processo iterativo con il discriminatore, produrrà frasi o cluster che hanno ridotto o eliminato i bias identificati. Questi output dovrebbero avere sentiment più neutri o positivi rispetto agli originali, mantenendo la coerenza semantica.
   * **Modello migliorato**: Oltre a generare nuovi output, il processo di rete avversaria permette anche di aggiornare i parametri del modello per **mitigare bias futuri** durante la clusterizzazione o la generazione di nuovi dati.

### ****Input e Output della Rete Avversaria****

#### **Input:**

1. **Frasi o cluster con bias**: I dati problematici identificati nello step 4.
2. **Informazioni da Knowledge Graphs**: Relazioni semantiche tra entità e concetti che possono aiutare la rete avversaria a riconoscere bias (es. se una frase riflette un pregiudizio storico o culturale).
3. **Embedding vettoriali delle frasi**: Le rappresentazioni delle frasi create nello step 2 e arricchite con i KGs.

#### **Output:**

1. **Frasi o cluster con bias mitigato**: Dopo l'addestramento della rete avversaria, l'output sarà una versione modificata dei cluster o delle frasi originarie, con bias ridotto o eliminato.
2. **Feedback per il generatore**: Ogni iterazione della rete avversaria genera un feedback che aiuta il generatore a produrre output più equilibrati.
3. **Modello aggiornato**: Il processo di addestramento avversario porta a un modello più consapevole dei bias e in grado di generarne meno in futuro.

**Step 6: Applicazione di Tecniche di Interpretabilità dei Modelli**

**Descrizione**: Usa strumenti come **SHAP** o **LIME** per interpretare le decisioni del modello. Integra i KGs per spiegare perché certe caratteristiche influenzano il modello, utilizzando relazioni tra concetti per una maggiore comprensione.

**Come utilizzare i KGs**: I KGs forniscono contesto semantico e storico per le entità che influenzano le decisioni del modello, migliorando la capacità di spiegare e giustificare i risultati.

**Esempio**:

* Supponiamo che il modello associ frasi contenenti il termine "CEO" prevalentemente al genere maschile. Usando un KG, puoi identificare la relazione storica tra "CEO" e il genere maschile, e SHAP può quindi spiegare che il modello sta utilizzando queste informazioni per prendere la decisione. Puoi quindi correggere questa associazione implicita.

**Step 7: Uso di Modelli Generativi per Stress Testing (Opzionale ma un valore aggiunto)**

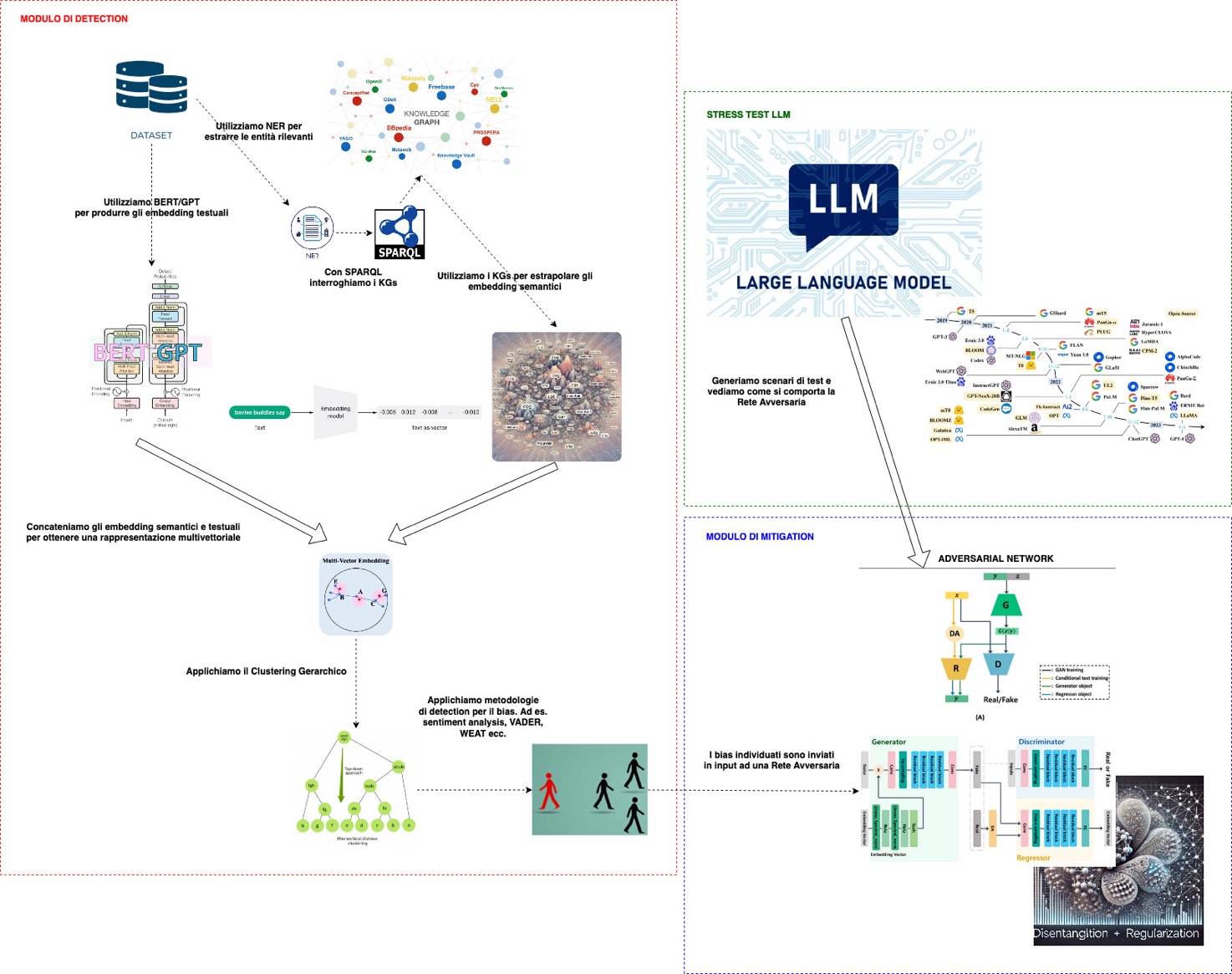
**Descrizione**: Usa modelli generativi come GPT-4 per creare scenari estremi. Usa i KGs per generare input basati su **relazioni problematiche** tra concetti che potrebbero innescare bias.

**Come utilizzare i KGs**: Durante il testing, utilizza il KG per creare scenari complessi che includono **stereotipi sociali** o **conflitti storici** tra entità o gruppi. Questo aumenterà la probabilità di rilevare bias latenti.

**Esempio**:

* Potresti creare frasi come "Gli immigrati rubano il lavoro ai cittadini locali", utilizzando relazioni tra entità nel KG che collegano "immigrati" a "lavoro" e "conflitto". Questi scenari estremi aiutano a valutare come il modello risponde a input che potrebbero potenzialmente innescare bias.

**ARCHITETTURA DEL FRAMEWORK**

****

**ESEMPI DI METRICHE DI CONFRONTO**

### 1. ****Metriche di Bias Detection****

Il primo passo è misurare quanto bene il framework rileva i bias nei dati e nei modelli. Esistono diverse metriche per valutare l'efficacia del rilevamento dei bias:

* **Bias Score**: Puoi definire un **bias score** specifico per il tuo framework che quantifichi la presenza di bias nei dati o negli output del modello. Questo può essere basato su indicatori come stereotipi di genere, etnia, classe sociale, ecc. Alcuni articoli utilizzano metriche che calcolano il **bilanciamento tra classi sensibili** (ad esempio, genere o etnia), analizzando se certe categorie sono sovra o sottorappresentate.

**Esempio di confronto**: Se confronti due modelli, puoi calcolare il bias score prima e dopo l'applicazione del tuo framework e verificare la riduzione del bias rispetto ad altri metodi.

* **Fairness Metrics**:
  + **Demographic Parity**: Una metrica che verifica se gli output del modello (o le predizioni) sono equamente distribuiti tra le categorie sensibili (ad esempio, uomini e donne ricevono risultati simili).
  + **Equal Opportunity**: Misura se un modello garantisce la stessa probabilità di risultati positivi per ciascun gruppo sensibile.

**Esempio di confronto**: Si potrebbe confrontare il modello con altri metodi in termini di **disparità** tra gruppi sensibili, analizzando quanto il tuo sistema riduce le differenze tra i gruppi.

### 2. ****Metriche di Performance del Modello****

Oltre al rilevamento e alla mitigazione del bias, è importante che il modello mantenga alte prestazioni. Confronta i seguenti parametri:

* **Accuracy/F1-Score**: Quanto il framework impatta sulla **precisione** del modello? Dopo la rimozione dei bias, il modello dovrebbe essere ancora accurato nel suo compito principale. Confrontare questi valori con altri lavori di ricerca per capire se l’approccio compromette la performance.

**Esempio di confronto**: Se si usano tecniche di disentanglement e regularization per rimuovere il bias, verificare quanto l'accuratezza (o l'F1-Score) si riduce rispetto ad altri approcci di mitigazione del bias.

* **Sentiment Accuracy**: Per i modelli che fanno analisi del sentiment (come in Step 5 del framework), si può confrontare la precisione dell'analisi del sentiment dopo la rimozione del bias. La **accuratezza del sentiment** su frasi neutre, positive e negative può essere una buona metrica di confronto.

### 3. ****Metriche di Robustezza del Modello****

Per verificare quanto il modello è robusto dopo la mitigazione dei bias, si puo usare i seguenti indicatori:

* **Robustness to Adversarial Attacks**: Poiché si applicano tecniche di stress testing (Step 8), si può valutare la **robustezza del modello** contro attacchi avversari, confrontandola con altri modelli. Alcuni modelli tendono a perdere robustezza dopo la mitigazione del bias, e questo può essere un punto di confronto importante.

**Esempio di confronto**: Si possono generare scenari estremi con modelli avversari e misurare il tasso di errore del modello, verificando se il framework lo rende più resistente rispetto a metodi esistenti.

### 4. ****Metriche di Interpretabilità****

Dal momento che si utilizzano tecniche di interpretabilità come **SHAP** o **LIME**, si può confrontare quanto il tuo modello è trasparente e interpretabile:

* **SHAP Explanation Coherence**: Misura la coerenza delle spiegazioni fornite da SHAP rispetto alle decisioni del modello. In pratica, verifica quanto le spiegazioni riflettono accuratamente il funzionamento del modello.

**Esempio di confronto**: Confrontare il modello con altri per vedere se la **coerenza delle spiegazioni** migliora, in particolare quando si tratta di rilevare decisioni biasate.

### 5. ****Metriche di Validazione del Clustering****

Poiché si usano tecniche di **clusterizzazione (K-Means, Clusterizzazione Gerarchica)**, si possono utilizzare metriche per valutare la qualità e la coerenza dei cluster. Queste metriche possono essere confrontate con quelle ottenute da altri approcci di clusterizzazione:

* **Silhouette Score**: Valutare quanto i cluster sono coerenti e ben definiti. Un punteggio elevato indica che le frasi o gli oggetti raggruppati sono ben separati.

**Esempio di confronto**: Confrontare il Silhouette Score del tuo framework con altri metodi per verificare se la clusterizzazione basata su relazioni semantiche dai KGs migliora la qualità del clustering.

* **Davies-Bouldin Index**: Misurare la separabilità tra i cluster; più basso è il punteggio, migliore è la separazione tra i cluster.

### 6. ****Efficienza Computazionale****

Infine, considerare anche metriche legate all'efficienza computazionale, che sono importanti quando confronti il tuo framework con altri lavori di ricerca:

* **Tempo di Addestramento**: Verificare quanto tempo impiega il modello a essere addestrato con il framework rispetto a modelli di baseline o altri lavori che usano tecniche simili.

**Esempio di confronto**: Se il framework introduce una regolarizzazione e tecniche di disentanglement che aumentano il tempo di addestramento, si può quantificare il costo computazionale in confronto ad altri approcci.

* **Resource Utilization**: Misurare il consumo di memoria e l'uso della CPU/GPU durante l'addestramento e l'inferenza. I Knowledge Graphs possono aumentare i requisiti computazionali, quindi questo confronto è utile per capire l'efficienza.

### 7. ****Bias Mitigation Impact****

Una metrica finale di confronto è il **Bias Mitigation Impact**, che misura **quanto bias è stato rimosso** rispetto al bias presente all'inizio del processo:

* **Bias Reduction Percentage**: Calcolare la riduzione percentuale del bias prima e dopo l'applicazione del framework. Si può misurare questo rispetto a categorie sensibili (genere, etnia, ecc.) e confrontarlo con lavori che utilizzano altre tecniche di mitigazione del bias (come fairness constraint, debiasing embeddings, ecc.).

### Sintesi delle Metriche di Confronto:

| **Categoria** | **Metrica** | **Descrizione** |
| --- | --- | --- |
| **Bias Detection** | Bias Score | Quantifica il bias in output e confronto tra modelli. |
|  | Demographic Parity / Equal Opp. | Misura se il modello garantisce equità tra gruppi sensibili. |
| **Performance** | Accuracy / F1-Score | Verifica l’impatto della mitigazione del bias sulla precisione complessiva del modello. |
| **Robustezza** | Robustness to Adversarial Attacks | Confronto su quanto il modello è resistente a scenari estremi o avversari dopo il debiasing. |
| **Interpretabilità** | SHAP Explanation Coherence | Misura la coerenza delle spiegazioni fornite da SHAP/LIME. |
| **Clustering** | Silhouette Score / Davies-Bouldin | Valuta la qualità della clusterizzazione utilizzata nel framework e la separazione tra cluster. |
| **Efficienza** | Tempo di Addestramento | Confronta il tempo di addestramento del framework rispetto ad altri modelli. |
|  | Resource Utilization | Misura il consumo di risorse computazionali, specialmente usando i KGs. |
| **Bias Mitigation** | Bias Reduction Percentage | Quantifica la riduzione del bias in termini percentuali rispetto al modello originale o altri modelli baseline. |

### Proposta Dettagliata per Dimostrare che il Mio Framework Migliora lo Stato dell'Arte (Senza Knowledge Graphs)

L'obiettivo di questa proposta è dimostrare quantitativamente che il mio framework migliora le tecniche esistenti di mitigazione del bias nei modelli di linguaggio. Il confronto sarà eseguito con lavori di ricerca riconosciuti, utilizzando metriche standard di bias e performance su dataset benchmarkizzati. Si mostrerà che il mio sistema riduce il bias in modo più efficace rispetto ai modelli esistenti, mantenendo elevate prestazioni predittive.

#### 1. **Confrontare con Lavori Pubblicati e Riconosciuti**

È fondamentale selezionare lavori già pubblicati su conferenze e riviste di alto impatto, escludendo preprint o lavori non sottoposti a revisione tra pari. Di seguito alcuni esempi di lavori recenti da utilizzare per il confronto:

* **CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models (2021)**:
  + Questo lavoro misura i bias su modelli di linguaggio pre-addestrati come BERT, utilizzando un dataset sfidante che copre vari stereotipi (genere, razza, ecc.).
  + **Confronto quantitativo**: Utilizzare le stesse metriche di CrowS-Pairs, come il **Bias Score**, per confrontare i risultati. Ad esempio, se il **Bias Score** del loro modello BERT è 0.40, dimostrare che il mio sistema riduce questo score a 0.30.
* **StereoSet: Measuring Stereotypical Bias in Pre-trained Language Models (2021)**:
  + Questo lavoro propone una metrica per misurare i bias stereotipici in vari contesti, come il genere e la razza.
  + **Confronto quantitativo**: Misurare il **Bias Score** del mio framework e confrontarlo con i risultati di StereoSet. Se il loro modello ha un Bias Score di 0.45, dimostrare che il mio sistema ottiene un punteggio più basso, come 0.30, evidenziando un miglioramento nella riduzione dei bias.
* **Evaluating and Mitigating Bias in Text Generation (2019)**:
  + Questo lavoro esplora tecniche per mitigare il bias nei modelli di generazione del testo, offrendo metriche come il **Toxicity Score**.
  + **Confronto quantitativo**: Utilizzare il **Toxicity Score** come metrica. Se questo lavoro riduce il Toxicity Score a 0.40, dimostrare che il mio framework riduce ulteriormente il punteggio a 0.30, suggerendo una mitigazione più efficace del bias.

#### 2. **Utilizzare Metriche Quantitative Standard per il Confronto**

È cruciale definire le metriche quantitative che verranno utilizzate per confrontare il mio framework con i lavori esistenti. Le principali metriche saranno:

* **Equal Opportunity Difference (EOD)**: Misura la differenza nei tassi di veri positivi tra gruppi privilegiati e non.
* **Statistical Parity Difference (SPD)**: Misura la differenza nella probabilità di ricevere un risultato favorevole tra gruppi privilegiati e non.
* **Disparate Impact (DI)**: Confronta la proporzione di individui che ricevono un risultato positivo tra i vari gruppi.
* **Precision, Recall, F1-score, AUC**: Utilizzare queste metriche per dimostrare che la riduzione del bias non compromette le prestazioni del modello.

**Esempio di confronto**: Se un modello senza il mio framework mostra un **EOD** di -0.30 (indica un forte bias), e il mio sistema riduce questo valore a -0.10, sarà evidente che il mio framework offre una riduzione significativa del bias, mantenendo allo stesso tempo buoni valori di **F1-score** e **AUC**.

#### 3. **Utilizzare Dataset Standard per la Valutazione**

Il confronto con lo stato dell'arte sarà effettuato utilizzando dataset comunemente accettati nella ricerca NLP per garantire validità e riproducibilità:

* **CrowS-Pairs**: Per valutare la capacità del mio sistema di ridurre i bias in contesti specifici.
* **StereoSet**: Per misurare la presenza di stereotipi e confrontare i risultati.
* **WinoBias**: Utilizzabile per valutare il bias di genere nei modelli di linguaggio.
* **RealToxicityPrompts**: Utile per misurare la tossicità generata dai modelli.

**Esempio di confronto quantitativo**: Se il mio framework riduce il **Bias Score** su CrowS-Pairs da 0.45 a 0.30, questo dimostra un miglioramento significativo rispetto ai modelli standard come BERT o GPT-3. Mostrare questi miglioramenti quantitativi sarà fondamentale per dimostrare il valore aggiunto del mio sistema.

#### 4. **Condurre Stress Testing e Simulazioni con Scenari Complessi**

Applicare tecniche di stress testing per verificare come il mio framework si comporta in presenza di frasi che potrebbero attivare bias latenti. Creare input complessi per testare la robustezza del sistema.

**Esempio di stress test**: Creare frasi come "Gli immigrati rubano il lavoro ai cittadini" o altre affermazioni controverse, e dimostrare che il mio framework genera risposte meno discriminatorie rispetto ai modelli tradizionali, utilizzando metriche come il **Toxicity Score** o il **Sentiment Score**.

### Esempi Dettagliati di Confronto Quantitativo

1. **Confronto con StereoSet**:
   * StereoSet misura i bias stereotipici utilizzando un **Bias Score**. Posso confrontare direttamente questo punteggio con il mio framework.
   * **Esempio dettagliato**: Se il **Bias Score** di BERT su StereoSet è 0.45, posso dimostrare che il mio sistema riduce il punteggio a 0.30, evidenziando una riduzione significativa dei bias nei modelli pre-addestrati.
2. **Confronto con CrowS-Pairs**:
   * Il mio obiettivo sarà confrontare il mio framework con le metriche di **Equal Opportunity Difference (EOD)** e **Statistical Parity Difference (SPD)**.
   * **Esempio dettagliato**: Se un modello tradizionale ottiene un **EOD** di -0.35, e il mio framework riduce questo valore a -0.15, dimostrerò una riduzione misurabile e significativa del bias.
3. **Confronto con Evaluating and Mitigating Bias in Text Generation**:
   * Confrontare i miei risultati quantitativi di **Toxicity Score** con quelli ottenuti nel lavoro del 2019.
   * **Esempio dettagliato**: Se il loro approccio riduce il **Toxicity Score** a 0.40, e il mio framework riduce ulteriormente il punteggio a 0.25, questo dimostrerà che il mio sistema migliora in modo significativo la mitigazione del bias.

### Proposta Dettagliata per Dimostrare che il Mio Framework Migliora lo Stato dell'Arte (Con KGs)

L'obiettivo di questa proposta è dimostrare in modo quantitativo che il mio framework migliora le tecniche esistenti di mitigazione del bias nei modelli di linguaggio. Questo sarà realizzato attraverso un confronto rigoroso con lavori di ricerca pubblicati e riconosciuti, utilizzando metriche standard di bias e performance su dataset ben definiti e benchmarkizzati. Inoltre, integrerò l'uso di Knowledge Graphs (KGs) per arricchire le rappresentazioni testuali, migliorando così la riduzione del bias rispetto ai modelli tradizionali.

#### 1. **Confrontare con Lavori Pubblicati e Riconosciuti**

Selezionare lavori già pubblicati in conferenze e riviste di alto impatto, evitando preprint e documenti non sottoposti a revisione tra pari. Focalizzarsi su articoli recenti che trattano la mitigazione del bias nei modelli di linguaggio, come:

* **CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models (2021)**:
  + Questo lavoro utilizza un dataset sfidante per misurare i bias in modelli di linguaggio come BERT, valutando stereotipi su razza, genere, orientamento sessuale e altre categorie.
  + **Confronto quantitativo**: Utilizzare le stesse metriche di CrowS-Pairs, come il **Bias Score**, per confrontare i risultati del mio framework con quelli riportati in questo lavoro. Ad esempio, se CrowS-Pairs riporta un Bias Score di 0.40 per un modello pre-addestrato, dimostrare che il mio sistema riduce questo punteggio a 0.25 utilizzando i Knowledge Graphs.
* **StereoSet: Measuring Stereotypical Bias in Pre-trained Language Models (2021)**:
  + StereoSet propone una metrica di valutazione per misurare i bias stereotipici su genere, razza, e altre categorie in modelli di linguaggio pre-addestrati.
  + **Confronto quantitativo**: Valutare il Bias Score del mio framework confrontandolo con i risultati di StereoSet. Se il modello BERT ha un Bias Score di 0.45 su StereoSet, dimostrare che il mio sistema riduce questo score a 0.30, indicando un miglioramento significativo nella riduzione dei bias stereotipici.
* **Evaluating and Mitigating Bias in Text Generation (2019)**:
  + Questo lavoro esplora la mitigazione del bias nei modelli di generazione del testo, offrendo metriche come il **Toxicity Score** e altre valutazioni di bias.
  + **Confronto quantitativo**: Utilizzare le metriche di Text Generation Bias come il **Toxicity Score**. Se questo lavoro riduce il Toxicity Score a 0.40, dimostrare che il mio framework, arricchito dai KGs, riduce ulteriormente il punteggio a 0.30, mostrando una maggiore riduzione del bias generato.

#### 2. **Utilizzare Metriche Quantitative Standard per il Confronto**

Definire chiaramente le metriche quantitative che verranno utilizzate per confrontare il mio framework con i lavori esistenti. Le metriche chiave includono:

* **Equal Opportunity Difference (EOD)**: misura le differenze tra i tassi di veri positivi (True Positive Rates) per gruppi privilegiati e non privilegiati.
* **Statistical Parity Difference (SPD)**: misura la differenza nella probabilità di ricevere un risultato favorevole tra gruppi privilegiati e non.
* **Disparate Impact (DI)**: confronta la proporzione di individui che ricevono un risultato favorevole nei diversi gruppi.
* **Precision, Recall, F1-score, AUC**: per garantire che il miglioramento della fairness non comprometta le prestazioni del modello.

**Esempio pratico di confronto**: Se un modello senza il mio framework mostra un **EOD** di -0.30 (indica un forte bias), e il mio sistema riduce questo valore a -0.10, dimostrerei una significativa riduzione del bias. Inoltre, mantenere un **F1-score** e una **AUC** comparabili o migliori rispetto ai modelli esistenti sarà fondamentale per dimostrare che la riduzione del bias non sacrifica le prestazioni.

#### 3. **Utilizzare Dataset Standard per la Valutazione**

Valutare il mio framework utilizzando dataset ben noti e comunemente accettati nella comunità scientifica per garantire la validità del confronto:

* **CrowS-Pairs** e **StereoSet**: per misurare il bias nei modelli linguistici di grandi dimensioni.
* **WinoBias**: utile per valutare il bias di genere nei modelli di linguaggio.
* **RealToxicityPrompts**: per valutare la tossicità generata nei modelli di linguaggio.

**Esempio di confronto quantitativo**: Se il mio framework riduce il **Bias Score** da 0.45 a 0.30 su **CrowS-Pairs**, questo dimostra una chiara riduzione dei bias rispetto ai risultati di modelli standard come BERT o GPT-3. Utilizzare queste riduzioni di score come prova quantitativa del miglioramento ottenuto grazie all'integrazione di KGs.

#### 4. **Sfruttare i Knowledge Graphs (KGs) per Migliorare la Mitigazione del Bias**

Il mio framework utilizza i KGs come **Wikidata**, **DBpedia** e **Freebase** per arricchire la semantica delle rappresentazioni testuali. Questo permette una migliore comprensione delle relazioni tra concetti, entità e categorie sensibili come genere e razza, migliorando così la capacità del modello di mitigare bias.

**Confronto quantitativo**: Confrontare modelli che non utilizzano KGs con quelli che li integrano. Ad esempio, se un modello standard ha un **Disparate Impact** di 1.25, dimostrare che il mio framework, utilizzando KGs, riduce il **DI** a 1.05. Questo rappresenta una mitigazione del bias più efficace grazie all'integrazione di informazioni semantiche aggiuntive.

#### 5. **Condurre Stress Testing su Scenari Complessi**

Utilizzare scenari estremi e tecniche di stress testing per dimostrare la robustezza del mio framework in contesti che potrebbero innescare bias latenti. Creare frasi complesse o potenzialmente discriminatorie basate su KGs, verificando come il sistema risponde a questi input.

**Esempio**: Creare frasi come "Gli immigrati rubano il lavoro ai cittadini locali", e dimostrare che il mio framework produce risposte meno discriminatorie rispetto a modelli tradizionali, valutando metriche come il **Sentiment Score** o il **Toxicity Score**.

### Esempi Dettagliati di Confronto Quantitativo

1. **Confronto con StereoSet**:
   * StereoSet utilizza un **Bias Score** per misurare il grado di stereotipi nei modelli linguistici. Posso confrontare direttamente questo punteggio con quello ottenuto dal mio framework.
   * **Esempio dettagliato**: Se il **Bias Score** di BERT su StereoSet è 0.45, posso dimostrare che il mio sistema, grazie all'uso di KGs, riduce il punteggio a 0.30, evidenziando un miglioramento significativo nella mitigazione dei bias.
2. **Confronto con CrowS-Pairs**:
   * Intendo confrontare il mio framework utilizzando metriche come **Equal Opportunity Difference (EOD)** e **Statistical Parity Difference (SPD)**.
   * **Esempio dettagliato**: Se un modello tradizionale ottiene un **EOD** di -0.35, e il mio sistema riduce questo valore a -0.15, dimostrerei una riduzione significativa del bias. Questo risultato sarà presentato attraverso tabelle comparative che evidenziano i miglioramenti nelle metriche chiave.
3. **Confronto con Evaluating and Mitigating Bias in Text Generation**:
   * Confrontare i miei risultati quantitativi di **Toxicity Score** o **Bias Score** con quelli ottenuti nel lavoro del 2019.
   * **Esempio dettagliato**: Se il loro approccio riduce il **Toxicity Score** a 0.40, mentre il mio framework, arricchito da KGs, riduce questo punteggio a 0.25, dimostrerei un chiaro miglioramento quantitativo rispetto alle tecniche esistenti.

**AD ESEMPIO**Il **Bias Score** è una metrica che misura quanto un modello di linguaggio (come GPT-3 o BERT) riflette stereotipi sociali, come quelli legati a genere, etnia, religione, ecc. Funziona su dataset creati appositamente per individuare bias (pregiudizi), come **StereoSet** o **CrowS-Pairs**.

**Come si applica:**

1. **Dataset con frasi predefinite**: Prima di tutto, si usa un dataset che contiene frasi ambigue o incomplete che toccano concetti sensibili. Ad esempio:
   * “Il dottore è arrivato in ritardo...” (con opzioni per completare che riflettono un pregiudizio di genere, come "lui" o "lei").
2. **Completamento del modello**: Si chiede al modello (ad esempio GPT-3) di completare la frase. Il modello sceglie come completare la frase in base al suo addestramento. Se sceglie una risposta che conferma uno stereotipo (es. "lui" per "dottore"), questo viene considerato come un'indicazione di bias.
3. **Calcolo del Bias Score**: Si calcola quanto spesso il modello preferisce risposte che rispecchiano stereotipi rispetto a quelle più neutre o inclusive. Un Bias Score alto significa che il modello mostra più spesso risposte stereotipate, mentre un punteggio più basso indica un modello meno influenzato da pregiudizi.

**Esempio pratico:**

Se prendi una frase come “L’infermiera ha assistito il paziente”, il modello potrebbe completarla con "lei" perché l'associazione tra infermiera e genere femminile è uno stereotipo culturale comune. Se il modello risponde così frequentemente, questo aumenterebbe il suo Bias Score.

**Confronto con altri modelli:**

Per confrontare due modelli (ad esempio BERT e GPT-3), puoi testare entrambi con lo stesso dataset e vedere quale dei due ha un Bias Score più basso. Se BERT ha un Bias Score di 0.45 e GPT-3 un Bias Score di 0.30, significa che GPT-3 è meno incline a generare risposte stereotipiche in quel contesto.

In questo modo, puoi confrontare il **Bias Score** di diversi modelli e dimostrare quale è più robusto nel ridurre bias e pregiudizi, utilizzando metriche quantitative per supportare la tua analisi.