**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.

**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.

**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.

**Step 3: Clusterizzazione Iniziale delle Rappresentazioni Testuali**

**Descrizione**: Applica algoritmi di clusterizzazione (ad es. **K-Means** o **Agglomerative Clustering**) per raggruppare le frasi in base alle rappresentazioni multivettoriali. Con l’integrazione dei KGs, i cluster non saranno basati solo sulla somiglianza testuale, ma anche sulle relazioni semantiche.

**Come utilizzare i KGs**: Usa le relazioni nei KG per **informare e migliorare il processo di clusterizzazione**. Ad esempio, potresti clusterizzare le frasi che contengono concetti simili o collegati tra loro all'interno del KG, indipendentemente dalla somiglianza lessicale.

**Esempio**:

* Immagina un cluster che contiene frasi come "La donna ha lasciato la sua carriera per prendersi cura dei figli" e "L'uomo è sempre al lavoro mentre la moglie resta a casa". Queste frasi, oltre a essere simili per significato, possono essere correlate tramite un KG che collega concetti come **"ruoli di genere tradizionali"**, **"divisione del lavoro"** e **"bias di genere nelle professioni"**.

**Step 4: Aggiunta di Tecniche di Disentanglement + Regularization**

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

**Come utilizzare i KGs**: I KGs ti 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.

**Step 5: Analisi Sentimentale sui Cluster**

**Descrizione**: Applica strumenti di **analisi del sentiment** come VADER o TextBlob sui cluster creati. Utilizza 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.

**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: 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.

**Step 8: Uso di Modelli Generativi per Stress Testing**

**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.

**Step 9: Clusterizzazione e Rilevamento del Bias a Livello Gerarchico**

**Descrizione**: Continua a iterare la clusterizzazione e il rilevamento del bias, utilizzando i KGs per identificare bias più complessi e contesti specifici.

**Come utilizzare i KGs**: I KGs possono fornire una guida semantica per esplorare nuove dimensioni di bias. Ogni iterazione del ciclo di clusterizzazione può sfruttare informazioni aggiornate dal KG.

**Esempio**:

* Se scopri che un cluster contiene frasi legate al genere, puoi usare il KG per espandere l'analisi includendo frasi che riguardano la discriminazione basata sull'**orientamento sessuale** o la **razza**, grazie alle relazioni semantiche nel grafico.

**Step 10: 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.

**DOVE TORNANO UTILI I KGs**

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

**Uso di KGs**: **Facoltativo**, ma utile. I KGs possono essere utilizzati per arricchire i dati, fornendo **informazioni strutturate** su entità e relazioni. Tuttavia, se hai già un dataset bilanciato e diversificato, l'uso dei KGs potrebbe essere meno cruciale.

* **Quando usarli**: Se hai bisogno di garantire che le entità all'interno dei tuoi dataset siano ben contestualizzate (ad es., identificare la provenienza storica o culturale delle professioni menzionate).

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

**Uso di KGs**: **Molto utile**. Integrare i KGs per creare vettori ibridi (combinando embedding generati dalle LLM con informazioni strutturate dai KGs) può migliorare la rappresentazione semantica delle frasi, arricchendo il contesto.

* **Quando usarli**: Quando vuoi migliorare la capacità del modello di comprendere relazioni tra concetti, specialmente quando le frasi fanno riferimento a entità o eventi storici, sociali o culturali.

**Step 3: Clusterizzazione Iniziale delle Rappresentazioni Testuali**

**Uso di KGs**: **Utile**. I KGs possono migliorare la clusterizzazione fornendo **relazioni semantiche** tra le frasi, consentendo di raggrupparle non solo in base alla somiglianza lessicale, ma anche in base al contesto concettuale.

* **Quando usarli**: Quando la clusterizzazione basata su dati puramente testuali potrebbe perdere collegamenti importanti tra entità legate da relazioni semantiche o culturali, che non sono evidenti a livello superficiale.

**Step 4: Aggiunta di Tecniche di Disentanglement + Regularization**

**Uso di KGs**: **Utile, ma non necessario**. I KGs possono aiutare a identificare e separare meglio gli attributi sensibili (come genere, etnia o professione) dalle rappresentazioni latenti, ma non sono obbligatori per il disentanglement, che può essere eseguito con tecniche avanzate di machine learning.

* **Quando usarli**: Se vuoi una comprensione più approfondita degli attributi sensibili, collegandoli a relazioni storiche o sociali specifiche.

**Step 5: Analisi Sentimentale sui Cluster**

**Uso di KGs**: **Facoltativo**. I KGs possono migliorare l'analisi del sentiment fornendo contesto alle frasi ambigue o aiutando a collegare entità e concetti che portano a sentimenti polarizzanti. Tuttavia, molti strumenti di sentiment analysis possono funzionare bene anche senza i KGs.

* **Quando usarli**: Se l'analisi del sentiment non riesce a rilevare con precisione bias impliciti o connotazioni negative che dipendono dal contesto culturale o sociale.

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

**Uso di KGs**: **Molto utile**. I KGs possono migliorare l'interpretabilità spiegando in modo esplicito perché il modello prende certe decisioni, collegando le entità e i concetti a contesti più ampi.

* **Quando usarli**: Quando vuoi comprendere e spiegare in modo più approfondito come il modello associa certe decisioni a determinati concetti o entità, soprattutto in caso di bias legati a relazioni storiche o culturali.

**Step 7: Clusterizzazione Gerarchica con Validazione**

**Uso di KGs**: **Facoltativo, ma utile**. I KGs possono supportare la clusterizzazione gerarchica fornendo informazioni semantiche che aiutano a validare la coerenza dei cluster, ma non sono strettamente necessari per le metriche standard di validazione (Silhouette Score, Davies-Bouldin Index).

* **Quando usarli**: Quando hai bisogno di un'ulteriore validazione semantica dei cluster, in particolare se i cluster identificano contesti sociali o culturali specifici che i KGs possono mappare.

**Step 8: Uso di Modelli Generativi per Stress Testing**

**Uso di KGs**: **Molto utile**. I KGs possono aiutare a creare scenari di **stress test** più ricchi e mirati, fornendo input che riflettono relazioni problematiche tra concetti sensibili o storicamente controversi.

* **Quando usarli**: Quando stai generando scenari estremi o complessi e vuoi assicurarti che i bias storici o culturali siano inclusi nei test per verificare la robustezza del modello.

**Step 9: Clusterizzazione e Rilevamento del Bias a Livello Gerarchico**

**Uso di KGs**: **Utile, ma non essenziale**. I KGs possono fornire una mappa delle relazioni tra concetti che permette di rilevare bias complessi e sottili nei cluster, ma non sono obbligatori in questa fase, che può essere gestita anche con i soli dati testuali.

* **Quando usarli**: Se il bias risulta particolarmente complesso o difficile da identificare, il KG può aiutare a collegare entità e concetti in modo più efficace, rivelando bias più nascosti.

**Step 10: Integrazione tra Clusterizzazione e Rete Avversaria**

**Uso di KGs**: **Molto utile**. I KGs possono aiutare le reti avversarie a identificare più facilmente pattern ricorrenti di bias, sfruttando le relazioni tra entità. Questo permette al sistema avversario di correggere bias in modo più informato.

* **Quando usarli**: Se vuoi che il sistema avversario possa apprendere dinamicamente dalle relazioni storiche o sociali presenti nei KGs, migliorando la capacità di rilevamento e mitigazione automatica del bias.

**Riepilogo: Quando usare i KGs**

1. **Obbligatorio/Molto utile**:
   * **Step 2 (Rappresentazione Multivettoriale)**: Per arricchire le rappresentazioni semantiche delle frasi.
   * **Step 6 (Interpretabilità dei Modelli)**: Per spiegare come e perché il modello prende certe decisioni.
   * **Step 8 (Stress Testing)**: Per generare scenari di test complessi e individuare bias nascosti.
   * **Step 10 (Clusterizzazione e Rete Avversaria)**: Per migliorare il rilevamento dinamico e la correzione del bias.
2. **Utile, ma non essenziale**:
   * **Step 3 (Clusterizzazione Iniziale)**: Per creare cluster semantici più precisi.
   * **Step 4 (Disentanglement e Regularization)**: Per separare meglio gli attributi sensibili.
   * **Step 7 (Clusterizzazione Gerarchica)**: Per validare ulteriormente i cluster a livello semantico.
   * **Step 9 (Rilevamento del Bias a Livello Gerarchico)**: Per identificare bias più complessi.
3. **Facoltativo**:
   * **Step 1 (Raccolta dei Dataset)**: Per arricchire il contesto delle entità nei dati.
   * **Step 5 (Analisi Sentimentale)**: Per migliorare la comprensione di sentiment ambigui o complessi.

**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**: Potresti confrontare il tuo 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 tuo framework impatta sulla **precisione** del modello? Dopo la rimozione dei bias, il modello dovrebbe essere ancora accurato nel suo compito principale. Confronta questi valori con altri lavori di ricerca per capire se il tuo approccio compromette la performance.

**Esempio di confronto**: Se hai applicato tecniche di disentanglement e regularization per rimuovere il bias, verifica 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 tuo framework), puoi 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, puoi usare i seguenti indicatori:

* **Robustness to Adversarial Attacks**: Poiché stai applicando tecniche di stress testing (Step 8), puoi 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**: Puoi generare scenari estremi con modelli avversari e misurare il tasso di errore del modello, verificando se il tuo framework lo rende più resistente rispetto a metodi esistenti.

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

Dal momento che utilizzi tecniche di interpretabilità come **SHAP** o **LIME**, puoi 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**: Confronta il tuo 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 usi tecniche di **clusterizzazione (K-Means, Clusterizzazione Gerarchica)**, puoi 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**: Valuta 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**: Confronta 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**: Misura la separabilità tra i cluster; più basso è il punteggio, migliore è la separazione tra i cluster.

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

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

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

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

* **Resource Utilization**: Misura 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**: Calcola la riduzione percentuale del bias prima e dopo l'applicazione del tuo framework. Puoi 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. |