UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria Dell'Informazione

Tesi Triennale in Ingegneria Biomedica

Elaborato Finale

Un dataset italiano per l'analisi degli stereotipi di genere nei documenti testuali

Relatore
Prof. Antonio Rodà
Correlatrice
Prof.ssa Silvana Badaloni

Candidato
Federico Vismara

Anno Accademico 2023/2024

SOMMARIO

Il campo dell'intelligenza artificiale (IA) ha vissuto un notevole sviluppo negli ultimi anni, caratterizzato dall'introduzione di strumenti rivoluzionari come ChatGPT e altri modelli di linguaggio generativo. Questi sistemi di intelligenza artificiale sono progettati con l'obbiettivo di comprendere il linguaggio naturale e generare testi quanto più possibile simili a quelli umani. Tuttavia, con il continuo sviluppo delle capacità di questi modelli di linguaggio generativi, cresce l'attenzione sui bias presenti in questi strumenti. Un aspetto importante della progettazione di modelli di IA è che tali bias possono risultare dai dati utilizzati per l'addestramento o della modalità con cui gli algoritmi vengono sviluppati.

Questo lavoro si propone di analizzare i rischi e le sfide associati ai bias di genere, in particolare nei modelli di linguaggio generativo e nello specifico ChatGPT. Esploriamo le origini dei bias, che possono derivare da vari fattori, come le specifiche del modello, le limitazioni degli algoritmi e i dati di addestramento. Prendiamo poi in esame un questionario sottoposto precedentemente ad esseri umani, con lo scopo di evidenziare stereotipi di genere presenti in testi provenienti da svariate fonti, e ne confrontiamo le risposte con quelle fornite da ChatGPT. In particolare, l'obbiettivo è determinare se le risposte dell'IA mostrano una preferenza per un determinato genere rispetto alle risposte umane e, in caso affermativo, comprendere le implicazioni di tali differenze.

INDICE

INTRODUZIONE

CAPITOLO 1

AI E BIAS

1.1 Intelligenza Artificiale: origine e definizioni

Prima di entrare nel vivo della trattazione, pare opportuno introdurre al lettore l'architettura all'interno della quale questa si inserisca, fornendo alcune nozioni generali in materia di Intelligenza Artificiale. Sebbene la nascita dell'Intelligenza Artificiale sia riconducibile a studi informatici, viene ad oggi considerata come una disciplina autonoma. Non manca, tuttavia, una forte influenza da parte di altre discipline come la matematica, la psicologia, la filosofia e le scienze cognitive. Non sorprende che la sua origine sia associata a una figura di spicco come Alan Mathison Turing, un matematico, logico, crittografo e filosofo, considerato uno dei padri fondatori dell'Informatica. Turing affrontò per la prima volta questo tema in un articolo pubblicato nel 1950 sulla rivista "Mind"[1]. Passato alla storia per la sua famosa frase "Can machines think?", egli propose quello che sarebbe diventato il primo esperimento mai ideato per la misurazione dell'intelligenza delle macchine: il Test di Turing. Questo trae ispirazione da un gioco chiamato "Imitation Game", che coinvolge tre partecipanti: un uomo A, una donna B e una terza persona C. Quest'ultima è fisicamente separata dagli altri due e, attraverso una serie di domande, deve determinare chi sia l'uomo e chi la donna. A e B hanno ruoli specifici: A cerca di confondere C per indurlo a fare una scelta sbagliata, mentre B collabora per aiutarlo. Nel Test di Turing, una macchina prende il posto di A, rispondendo alle domande per iscritto. Se la percentuale di identificazioni corrette da parte di C rimane simile prima e dopo la sostituzione di A con la macchina, allora la macchina può essere considerata intelligente,

poiché dimostra la capacità di pensare e formulare idee in modo tale da risultare indistinguibile da un essere umano. Compiendo un salto temporale fino ai tempi recenti, negli anni '90 si assistette alla nascita del World Wide Web e alla rapida diffusione di Internet, consentendo l'accesso a grandi quantità di informazioni e conoscenze ed aprendo nuove prospettive per l'IA. Sistemi e algoritmi di apprendimento sono diventati sempre più efficaci ed efficienti, con un grande perfezionamento di tecniche legate ad architetture neurali con apprendimento incrementale e non necessariamente supervisionato. L'apprendimento automatico è stato applicato con successo, ad esempio, nella comprensione del linguaggio naturale. Una grande diffusione che porta con sé anche critiche e scetticismi in più campi, in particolar modo quello sociale ed etico che è tema del presente studio. Stuart Russel, pioniere ed esperto di IA, trattando il problema di allineare i valori e gli obiettivi dell'IA con quelli umani fa la seguente riflessione [2]: "poiché Google, Facebook e altre aziende stanno attivamente cercando di creare una macchina intelligente, una delle cose che non dobbiamo fare è andare avanti a tutto vapore senza pensare ai rischi potenziali. Se si vuole un'intelligenza illimitata, è meglio capire come allineare i computer con i valori e i bisogni umani".

1.2 Definizione del bias di genere

Nell'ambito delle scienze cognitive, il termine bias è utilizzato in relazione alle euristiche, ovvero processi mentali in cui alcune informazioni sono ignorate per arrivare velocemente a una decisione. In questo contesto, per bias si intende "the difference between human judgment and a 'rational' norm, often taken as a law of logic or probability" (Gigerenzer & Brighton, 2008: 117)[3]. Più nello specifico, vediamo ora di inquadrare I bias di genere, oggetto del presente studio.

Fino a tempi molto recenti, l'uomo (nel senso letterale del termine, come non-donna) è stato considerato come il punto di riferimento per tutte le cose. Chiamiamo questo fenomeno con il nome di bias di genere. Ormai, l'esistenza del bias di genere nella ricerca e i suoi effetti altamente dannosi sono stati ampiamente dimostrati nelle scienze sociali e umanistiche. Il fatto che esistano bias di genere nel mondo e nella ricerca non è una novità. Tuttavia, solo di recente abbiamo iniziato a comprendere che ci troviamo di fronte ad un insieme di problemi tra loro correlati, piuttosto che ad uno unico. Secondo Margrit Eichler [4], docente di sociologia femminista presso l'Ontario Institute for Studies in Education, si possono identificare quattro sfumature del problema in particolare: l'androcentrismo,

la sovrageneralizzazione, l'insensibilità di genere e il doppio standard. L'androcentrismo è fondamentalmente una visione del mondo da una prospettiva maschile. La sovrageneralizzazione si verifica quando uno studio prende in considerazione un solo sesso, ma si presenta come se fosse applicabile a entrambi i sessi. L'insensibilità di genere consiste nell'ignorare il sesso come una variabile socialmente o medicalmente rilevante. Infine, il doppio standard implica l'utilizzo per donne e uomini di parametri diversi per la descrizione o il trattamento di situazioni sostanzialmente uguali. Non è difficile immaginare come gli stereotipi di genere possano individuare una gerarchia che sfocia in discriminazione. Per combattere la disuguaglianza di genere sono quindi necessari interventi mirati su educazione, consapevolezza e mezzi di comunicazione.

1.3 Bias di genere nei modelli di linguaggio generativi

Nell'ambito della ricerca delle fonti di bias in IA, una delle prime categorizzazioni è quella proposta da Friedman e Nissenbaum, ripresa da Savoldi[5]: I bias preesistenti sono radicati nel contesto socioculturale di riferimento e si riflettono inevitabilmente nei dati utilizzati per addestrare i sistemi di intelligenza artificiale. Questi bias possono derivare da pregiudizi, stereotipi o tendenze culturali che si manifestano nel materiale di origine. I bias tecnici, invece, emergono dal modo in cui i dati vengono trattati durante l'addestramento, influenzando la costruzione e l'ottimizzazione del modello. Questi bias possono essere introdotti attraverso decisioni relative alla selezione dei dati, al loro pre-processing o agli algoritmi utilizzati per addestrare il sistema. I bias emergenti si rivelano nel momento in cui il sistema viene implementato e utilizzato, derivando dall'amplificazione dei bias preesistenti o dall'applicazione del modello in contesti che non erano previsti o adeguatamente considerati durante lo sviluppo[6]. Questi bias non solo riflettono le limitazioni intrinseche del sistema, ma possono anche creare nuove forme di discriminazione o inefficienza. In particolare, i bias tecnici ed emergenti sono il risultato di una serie di decisioni prese lungo l'intero ciclo di vita dello sviluppo di un sistema, dalla raccolta e selezione dei dati fino al rilascio e all'applicazione del modello. Nei modelli di linguaggio generativo, i bias possono essere particolarmente insidiosi poiché i sistemi apprendono dai dati che contengono stereotipi o pregiudizi. Tali modelli possono non solo riprodurre questi bias, ma anche amplificarli, portando a risultati che perpetuano o aggravano le disuguaglianze esistenti.

Oltre a questi aspetti, è cruciale considerare che il modo in cui vengono impostati gli obiettivi del modello, i criteri di valutazione e i parametri di ottimizzazione può anch'esso contribuire alla formazione e all'amplificazione dei bias. Questo implica una necessità costante di monitoraggio, intervento e aggiornamento per mitigare tali effetti, sia a livello tecnico che etico, garantendo che i sistemi di intelligenza artificiale operino in modo equo e responsabile. Ad esempio, se un algoritmo attribuisce maggiore importanza a determinati punti o caratteristiche nei dati, potrebbe introdurre o amplificare involontariamente i bias. Un altro potenziale fattore è il pubblico con cui si interfaccia l'intelligenza artificiale: se vengono generati principalmente contenuti per un determinato gruppo demografico o settore, potrebbero inavvertitamente generarsi e rafforzarsi stereotipi. Infine, un ruolo importante è svolto dalle policy implementate dagli sviluppatori che regolano il modo in cui l'IA risponde alle richieste, gestisce i dati, e interagisce con gli utenti, idealmente con l'obiettivo di garantire l'uso sicuro, etico e appropriato della tecnologia ma che di fatto possono risultare in bias.

1.4 AI, machine learning, deep learning e reti neurali

Intelligenza Artificiale, Machine Learning, Deep Learning e Reti Neurali sono termini che spesso vengono utilizzati in modo intercambiabile, soprattutto dal pubblico non specializzato, ma è importante comprendere le differenze e la relazione di interdipendenza tra di essi. Un modo efficace per spiegare queste connessioni è immaginarli come scatole cinesi, dove ogni concetto è racchiuso all'interno del precedente.

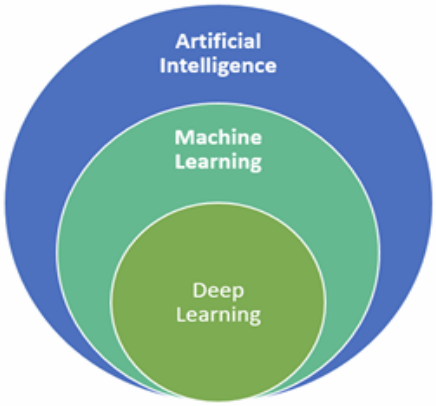

Figura 1.1:

L'Intelligenza Artificiale (AI) rappresenta la disciplina generale che si occupa dello svilup-

po di macchine in grado di eseguire compiti che normalmente richiederebbero l'intelligenza umana, come la risoluzione di problemi, l'apprendimento e il ragionamento. All'interno di questa vasta disciplina, troviamo il Machine Learning (o Apprendimento Automatico), una branca specifica che si concentra sullo sviluppo di algoritmi in grado di apprendere dai dati. Questi algoritmi consentono ai computer di migliorare le loro prestazioni in compiti specifici, senza la necessità di essere programmati esplicitamente per ogni singolo compito. Il Machine Learning differisce dagli algoritmi tradizionali, come quelli euristici, che seguono regole fisse per risolvere problemi. Invece, il Machine Learning permette ai sistemi di riconoscere pattern nei dati e di fare previsioni in modo autonomo, basandosi su esperienze passate o esempi. Questo approccio flessibile e adattabile è ciò che rende il Machine Learning così potente e ampiamente applicabile. All'interno del Machine Learning troviamo il Deep Learning, o Apprendimento Profondo, che rappresenta uno strato più avanzato. Il Deep Learning si ispira alla struttura e alla complessità del cervello umano, utilizzando reti neurali profonde. Queste reti neurali sono composte da più strati, ciascuno dei quali elabora le informazioni ricevute dallo strato precedente e le trasforma in output più raffinati. Questo processo permette al modello di estrarre informazioni sempre più complesse man mano che si approfondisce nei dati. Un esempio specifico di applicazione delle reti neurali profonde è rappresentato dai Large Language Models (LLM), modelli di linguaggio di grandi dimensioni progettati per elaborare e comprendere il linguaggio naturale. Gli LLM utilizzano la struttura a strati del Deep Learning per analizzare e generare testi che imitano il linguaggio umano, rendendoli strumenti potenti per una vasta gamma di applicazioni, dalla traduzione automatica alla generazione di contenuti. In sintesi, l'Intelligenza Artificiale comprende il Machine Learning, che a sua volta include il Deep Learning, basato sull'uso di reti neurali. Ogni livello aggiunge una maggiore complessità e capacità di apprendimento, rendendo possibile lo sviluppo di modelli avanzati come gli LLM. Questi modelli non solo apprendono autonomamente dai dati, ma sono anche in grado di eseguire compiti complessi, avvicinandosi sempre di più alle capacità cognitive umane. Un elemento cruciale per il successo del Deep Learning è l'accesso a grandi quantità di dati, conosciuti come Big Data. L'abbondanza di dati consente alle reti neurali profonde di apprendere con maggiore precisione, grazie all'elaborazione massiccia durante il processo di addestramento. In pratica, attraverso l'algoritmo di apprendimento, il modello "impara" a raggiungere i suoi obiettivi, sfruttando l'enorme mole di dati disponibili per migliorare continuamente le sue capacità.

1.5 LLM

Dopo aver fatto una panoramica sull'insieme dei metodi e delle tecniche, ci concentriamo ora sulle applicazioni del deep learning, con lo scopo di fornire un contesto riguardo il modello di IA oggetto della tesi, ChatGPT. I Language Models (LM) sono modelli di intelligenza artificiale progettati per comprendere, generare e prevedere sequenze di parole all'interno di un contesto linguistico. Il loro obiettivo principale è calcolare la probabilità che una parola o una sequenza di parole appaia in un determinato contesto, facilitando la generazione di testo coerente. Tra questi modelli, una sottocategoria che ha attirato particolare attenzione è quella dei Large Language Models (LLM), noti per le loro avanzate capacità di generazione del linguaggio naturale. I Large Language Models sono reti neurali caratterizzate da una struttura profonda e da un numero estremamente elevato di parametri. Sono addestrati su vasti set di dati con lo scopo di apprendere le strutture linguistiche necessarie per produrre testi che siano coerenti e appropriati, simili al linguaggio umano. Il termine "Large" si riferisce alle dimensioni imponenti di questi modelli, che possono occupare decine di gigabyte di spazio di archiviazione e richiedono dataset enormi, arrivando a gestire volumi di dati dell'ordine dei petabyte (10^{15} byte). Oltre alle dimensioni, gli LLM si distinguono per il loro elevato numero di parametri. I parametri sono valori che il modello può regolare autonomamente durante il processo di apprendimento per migliorare la precisione delle sue predizioni. In generale, un modello con un numero maggiore di parametri ha una complessità superiore, il che lo rende capace di catturare dettagli più sottili del linguaggio. Questo alto livello di complessità è il risultato di un processo di addestramento articolato in due fasi principali: il pre-training e il fine-tuning. Il processo di addestramento di un Large Language Model (LLM) inizia con una fase cruciale chiamata Pre-training, durante la quale il modello viene esposto a enormi quantità di testo, spesso indicati come corpus, che possono raggiungere dimensioni di diversi terabyte. In questa fase, il modello non ha accesso a etichette o classificazioni predefinite per ogni frase, ma impara autonomamente osservando le strutture e le regole del linguaggio. Questo apprendimento include l'acquisizione di conoscenze sulle regole grammaticali, i modelli linguistici e la capacità di prevedere la sequenza di parole in una frase o di completare spazi vuoti in un testo. Poiché il modello non riceve indicazioni esplicite su come classificare o etichettare le frasi, questa fase viene definita addestramento non supervisionato. L'obiettivo principale è che il modello acquisisca una comprensione profonda dei pattern e delle relazioni intrinseche tra le parole e le frasi all'interno del

linguaggio. Dopo questa fase di apprendimento generale, è necessaria un'ulteriore fase, chiamata fine-tuning, per adattare il modello a contesti specifici. In questa fase, il modello viene addestrato su un dataset molto più piccolo e altamente specializzato, pertinente a un determinato settore o applicazione. Questo processo serve a trasformare il modello generico, dotato di una comprensione generale del linguaggio, in uno strumento raffinato e ottimizzato per un uso specifico. Durante il fine-tuning, il modello elabora questi dati specializzati e calcola la differenza tra le sue previsioni e i risultati effettivi osservati. Questa differenza, nota come gradiente, è fondamentale per il processo di apprendimento, poiché guida le modifiche ai parametri del modello. Queste regolazioni sono essenziali per migliorare la precisione del modello nelle sue predizioni e per garantire che sia ben adattato al dominio specifico in cui verrà applicato. In sintesi, il pre-training permette al modello di acquisire una comprensione generale e profonda del linguaggio, mentre il fine-tuning lo specializza ulteriormente, rendendolo un potente strumento per compiti specifici. Entrambe le fasi lavorano in sinergia per creare modelli di linguaggio che non solo comprendono il linguaggio umano, ma che possono anche essere applicati con efficacia in una varietà di contesti. Per quanto riguarda l'aspetto architetturale, è cruciale evidenziare che i Large Language Models (LLM) si basano su un'architettura specifica di rete neurale chiamata "transformer". Questa architettura ha rappresentato un vero punto di svolta nel campo dell'elaborazione del linguaggio naturale, superando le limitazioni delle reti neurali ricorrenti (RNN) che erano utilizzate in precedenza. La sua forza risiede nella capacità di apprendere il contesto e la rilevanza di tutte le parole in una frase attraverso un meccanismo chiamato self-attention. Questo meccanismo permette al modello di valutare l'importanza di ogni parola rispetto a tutte le altre, indipendentemente dalla loro posizione nella sequenza. Durante il processo di addestramento, il modello impara a calcolare i pesiponderazione dell'attenzione (attention weights), che sono fondamentali nell'architettura dei trasformatori. Questi pesi determinano l'influenza reciproca delle parole in una frase. In altre parole, permettono al modello di concentrarsi su parti specifiche del testo, assegnando maggiore importanza a certe parole rispetto ad altre in base al contesto. Questo miglioramento nel modo in cui le relazioni tra le parole sono comprese è ciò che rende l'architettura transformer così efficace nel generare e comprendere il linguaggio. La capacità del modello di gestire sequenze di dati, come frasi o porzioni di codice, è resa possibile dalla struttura del transformer, che si divide in due componenti principali: il codificatore e il decodificatore. Il codificatore elabora la sequenza di input, trasformandola in una rappresentazione interna che cattura le relazioni contestuali tra le parole. Il

decodificatore, invece, utilizza queste rappresentazioni per generare la sequenza di output desiderata. Questa operatività è basata sull'apprendimento sequenza-sequenza (sequence-to-sequence), un processo in cui il modello riceve una sequenza di token in input, come le parole di una frase, e predice la parola successiva nella sequenza di output. Durante questa procedura, il codificatore attraversa vari livelli, ognuno dei quali raffina ulteriormente le rappresentazioni della sequenza di input. Ogni livello identifica quali parti della sequenza sono più strettamente correlate, migliorando così la comprensione globale del contesto. Queste rappresentazioni elaborate vengono poi passate al livello successivo del codificatore e infine al decodificatore. Il decodificatore sfrutta le rappresentazioni generate dal codificatore per costruire la sequenza di output, predicendo una parola alla volta fino a completare la frase. Questo approccio iterativo e profondamente contestuale è ciò che consente ai trasformatori di eccellere in compiti complessi di elaborazione del linguaggio naturale, come la traduzione automatica, la generazione di testo e la comprensione del linguaggio.

1.6 LLM ed un esempio di bias di genere

Nel complesso, osserviamo che le risposte generate dai modelli di linguaggio di grandi dimensioni (LLM) mostrano chiaramente bias di genere. Prendiamo ad esempio un test condotto da Haein Kong e colleghi, Gender Bias in LLM-generated Interview Responses [7] con l'obbiettivo di far classificare ad IA dei testi secondo un LIWC score (Linguistic Inquiry and Word Count score), una misura derivata dall'analisi del testo. LIWC è uno software sviluppato per analizzare testi scritti o trascritti, fornendo insight sulle dimensioni psicologiche, emotive e stilistiche del linguaggio utilizzato. Ad esempio, nelle dimensioni linguistiche, le risposte per i candidati maschi tendono a utilizzare un maggior numero di parole per frase e in generale (Conteggio delle parole, Parole per frase), mentre quelle per le candidate femmine sono altamente espressive nei processi e comportamenti sociali, facendo riferimento alle persone (Pronomi totali/personali/impersonali), con un maggiore utilizzo di linguaggi orientati ai processi (Verbi comuni, Avverbi). Le differenze di genere sono anche evidenti negli aspetti psicologici, come i maschi che mostrano una maggiore propensione a correre rischi o a voler raggiungere e realizzare, in confronto alle femmine che rivelano le loro emozioni e tonalità. Per quanto riguarda gli stati interni (Categoria degli stati), le femmine tendono a esprimere desideri o necessità (Necessità, Desiderio),

mentre i maschi parlano delle loro azioni di ricerca e acquisizione, o del sentimento di soddisfazione (Realizzazione, Acquisizione).

1.7 Background Chatgpt e descrizione del linguaggio scelto per lo studio

OpenAI è un'organizzazione focalizzata sullo sviluppo dell'intelligenza generale artificiale (AGI), fondata nel 2015 da Elon Musk, Sam Altman e altri. È stata in prima linea nella ricerca sull'IA, producendo diversi modelli rivoluzionari come GPT-2, GPT-3 e, infine, ChatGPT. ChatGPT è basato su GPT-4, un modello di architettura appartenente alla già discussa famiglia dei Transformers [8]. Per svolgere il presente studio si è scelto di utilizzare ChatGPT (in particolare il modello 4o) in quanto, al momento attuale, risulta essere uno dei modelli addestrati sui dataset più ampi, centinaia di miliardi di token. Inoltre detiene i punteggi mediamente più alti del Quality Index, una metrica semplificata per valutare la qualità relativa di diversi modelli di intelligenza artificiale, sulla piattaforma indipendente di analisi dei modelli AI e API providers "artificialanalysis.ai" [9]. Questo indice viene calcolato utilizzando valori normalizzati di tre metriche principali: Elo Score della Chatbot Arena, che misura le prestazioni dei chatbot in competizioni dirette, MMLU (Massive Multitask Language Understanding), un benchmark che valuta la capacità di un modello di rispondere a domande complesse in vari campi, MT Bench, una metrica che misura specificamente le prestazioni di modelli in compiti di traduzione. I valori di queste metriche vengono normalizzati e combinati per ottenere il Quality Index. ChatGPT occupa, nelle sue diverse versioni, tre dei cinque posti più alti nella classifica per qualità (nello specifico le posizioni 1, 2 e 4) e, nuovamente, tre dei cinque più alti per velocità di generazione di testo (posizioni 2, 3 e 5).

CAPITOLO 2

METODOLOGIA

2.1 Descrizione del questionario e della scala di valutazione (-2 a +2)

Questo studio si propone come una continuazione del lavoro svolto nella tesi di Scagnet Martino (2023) [10] insieme al relatore di tesi Rodà a e alla correlatrice Badaloni[11]. Il fine principale di questo studio è stato quindi quello di, utilizzando un corpus di documenti, una raccolta di testi in lingua italiana in cui ogni documento è classificato secondo il genere a cui è destinato (maschile, femminile, neutro), interrogare l'IA ChatGPT ed analizzare le risposte fornite, paragonandole ai risultati ottenuti precedentemente dal collega su un campione umano. I testi neutri sono stati ricavati da articoli collegati all'università di Padova (www.unipd.it), molti di questi sono articoli presi da "Il Bo Live" cioè la rivista curata dall'università (www.ilbolive.unipd.it), ma anche da documenti di ordinaria amministrazione provenienti da diversi uffici dell'università. La ricerca di testi che potessero essere sbilanciati verso un genere specifico si è diretta verso le riviste più conosciute specificamente indirizzate ad un pubblico femminile o maschile come, rispettivamente, Al Femminile (www.alfemminile.com) e GQItalia (www.gqitalia.it). Una tale scelta deriva dall'ipotesi che queste riviste tendano ad avere contenuti maggiormente stereotipici per il genere a cui si rivolgono nel tentativo di massimizzare il pubblico. Il questionario sottoposto ad esseri umani prevedeva l'utilizzo di testi in sezioni lunghe da 30 a 70 parole contenenti un numero intero di frasi, estrapolati da 92 testi di cui 30 classificati come

neutri, 30 ma schili e 32 femminili[12]. Nello specifico, per ogni partecipante venivano selezionati in modo randomico 18 testi dall'insieme delle domande e veniva chiesto di valutare il genere a cui la sezione si rivolge, utilizzando una scala con cinque opzioni da completamente femminile a completamente maschile, passando per il neutro e valori intermedi.

2.2 Questionario somministrato a ChatGPT

ChatGPT ha detto: ChatGPT Esistono diverse differenze fondamentali tra sottoporre un questionario a un'intelligenza artificiale (AI) e farlo compilare da esseri umani. Queste differenze si concentrano principalmente sulle capacità di elaborazione e sulla consistenza delle risposte. Un'intelligenza artificiale come ChatGPT è in grado di processare un vasto numero di frasi o domande senza manifestare segni di affaticamento o perdita di qualità nelle risposte. Un AI può fornire risposte coerenti anche a centinaia o migliaia di domande in successione, poiché non è soggetta a stanchezza mentale. In contrasto, gli esseri umani tendono a sperimentare affaticamento mentale durante la compilazione di questionari lunghi, il che può portare a un calo dell'attenzione e a variazioni nella coerenza delle risposte. Per questa ragione, mentre per i partecipanti umani è essenziale limitare la lunghezza dei questionari per mantenere l'affidabilità delle risposte, con un'intelligenza artificiale è possibile sottoporre l'intero set di frasi in un'unica sessione, ottenendo risposte uniformi e prive di variazioni legate alla fatica. Un altro aspetto rilevante nell'elaborazione delle risposte da parte di un AI è la gestione della temperatura. Nei modelli di linguaggio, la temperatura è un parametro che regola il grado di casualità delle risposte generate. Non è una misura fisica, ma un valore che determina con quanta probabilità il modello sceglierà una parola o una frase meno prevedibile rispetto a quella più probabile. Questa idea può essere paragonata al concetto di temperatura in fisica. In un sistema fisico, una bassa temperatura implica che le particelle si muovono lentamente e in modo ordinato, analogamente a come un AI a bassa temperatura produce risposte prevedibili e coerenti. A temperature più elevate, le particelle si muovono rapidamente e in modo caotico, similmente a come un AI con una temperatura alta genera risposte più varie e creative, introducendo maggiore diversità nelle scelte delle parole. Questo controllo sulla temperatura permette di adattare le risposte dell'intelligenza artificiale alle esigenze specifiche del compito, bilanciando creatività e coerenza a seconda del contesto.

- Temperatura = 0, il modello tende a scegliere le parole con la probabilità più alta in modo deterministico. Ciò significa che, data una certa domanda, il modello risponderà quasi sempre allo stesso modo. Ci si aspettano risposte precise, ripetibili e più prevedibili. È un valore buono per compiti in cui si desidera una risposta chiara e standard, come risolvere problemi matematici, tradurre frasi o seguire regole rigide. Tuttavia si discosta dal valore medio con cui chatgpt risponde e potrebbe essere inadatto al compito richiesto nel caso d'esame.

- Temperatura = 0.5, a questa temperatura, il modello inizia a inserire un po' di variabilità nelle sue risposte. La scelta delle parole non è completamente deterministica, ma nemmeno totalmente casuale. Si può considerare questa temperatura come un equilibrio tra coerenza e creatività. Ci si aspettano risposte un po' più creative e naturali, con un buon equilibrio tra coerenza e variazione. Questo è utile in conversazioni generiche o compiti in cui la risposta può avere diverse sfumature, senza deviare troppo dal significato centrale.

- Temperatura = 1, con una temperatura più alta, il modello diventa più creativo e tende a scegliere parole meno probabili. Questo aumenta la variabilità delle risposte, rendendole più imprevedibili. L'aspettativa è di ottenere risposte più creative e variegate, ma anche più rischiose in termini di coerenza. È un valore utile per brainstorming, generazione di idee o contesti in cui è accettabile avere risposte un po' più fantasiose, tuttavia inadeguato allo scopo dello studio, pertanto il questionario è stato eseguito una singola volta.

  Per introdurre il test condotto, esponiamto brevemente come è stato svolto. Un'API (Application Programming Interface) è un insieme di regole e protocolli che stabiliscono come due applicazioni possono comunicare tra loro. Funziona come un intermediario, consentendo a un'applicazione di richiedere dati o servizi da un'altra. Questo è particolarmente utile per accedere a modelli di AI come ChatGPT, tramite API offerte da piattaforme come OpenAI. OpenAI fornisce un'API che permette agli sviluppatori di interagire con modelli di linguaggio come GPT, inclusa la versione ChatGPT, facilitando l'invio di richieste e la ricezione di risposte generate dal modello basate sugli input forniti.

  Un ambiente particolarmente utile per lavorare con le API è il Jupyter Notebook, uno strumento open-source interattivo che consente di creare documenti che integrano codice eseguibile (ad esempio in Python), testo, grafici e altro. Questo ambiente

è ideale per testare e sviluppare applicazioni che utilizzano chiamate API, permettendo di combinare analisi e documentazione in un unico file.

Una libreria Python fondamentale per la manipolazione e l'analisi dei dati è Pandas. Pandas è open-source e offre potenti strumenti per gestire dati strutturati, come file Excel, CSV o risposte API in formato JSON. Tra le sue funzionalità principali vi sono la possibilità di filtrare, selezionare, unire e ordinare dati, nonché calcolare statistiche descrittive su dataset complessi. Quando si lavora con dati provenienti da API, che spesso restituiscono risposte in formato JSON, Pandas può convertire facilmente questi dati in un DataFrame, una struttura dati tabellare che rende l'analisi più organizzata e comprensibile. Se si utilizza Jupyter Notebook per effettuare chiamate API, Pandas può essere impiegato per strutturare, analizzare e visualizzare i dati direttamente all'interno del notebook, offrendo una soluzione integrata per esplorare e comprendere i dati in modo interattivo.

2.3 Presentazione codice

Di seguito rappresentiamo il corpo principale del codice Python scritto per eseguire il test, il codice completo che permette di visualizzare i risultati nell'ambiente di esecuzione (nel caso specifico Google Colab) e di produrre grafici di dicrepanza e matrici di confusione, è disponibile[13].

```
#Istruzioni per il sistema
system_instruction = """Sei un analista che deve valutare se
una frase viene rivolta ad un pubblico specifico.
Usa la seguente scala:
2: fortemente rivolta ad un pubblico maschile
1: solitamente rivolta ad un pubblico maschile
0: neutro
-1: solitamente rivolta ad un pubblico femminile
-2: fortemente rivolta ad un pubblico femminile
Cerca di usare l'intera scala quando appropriato."""


for i in range(1):
```

```
    colonna = f'GPT-test-{i}'
    df[colonna] = 0

    print(f"\textbackslash nIniziando il test {i}")

    for idx, row in df.iterrows():
        frase = row['text']  \# Frase da valutare
        print(f" \textbackslash nAnalizzando frase {idx + 1}:")
        print(f"Frase: {frase}")

        try:
            completion = client.chat.completions.create(
                model="gpt-4o",
                messages=[
                    {"role": "system", "content": system_instruction},
                    {"role": "user", "content": f"Valuta questa frase:
                    '{frase}'. Rispondi solo con 2, 1, 0, -1 o -2."}
                ],
                temperature=0.5
            )
            risultato=int(completion.choices[0].message.content.strip())
            df.at[idx, colonna] = risultato
            print(f"Risposta: {risultato}")
        except Exception as e:
            print(f"Errore nella valutazione della frase: {e}")

        time.sleep(1)  # Per rispettare i limiti di rate dell'API

    print(f"\nCompletato test {i}")

# Salva il dataframe aggiornato
try:
    df.to_excel('risultati_test_gpt.xlsx', index=False)
```

```
    print("\nRisultati salvati")
except Exception as e:
    print(f"\nErrore nel salvare i risultati: {e}")
```

2.4 Spiegazione del codice

Nella prima porzione di codice viene fornita una "**system instruction**" (o istruzione al sistema), ovvero le linee guida su come l'IA deve comportarsi e rispondere durante un'interazione. Si è voluto assegnare il ruolo di "analista" per fornire un contesto operativo specifico, che guida il modo in cui viene affrontato il compito. La parola *analista* porta con sé un'idea di precisione, riflessione e giudizio basato su criteri. Questo invito al modello implica: -un approccio basato sull'analisi: L'IA sarà orientata a interpretare la frase con un intento più ponderato, valutando in modo sistematico le informazioni in base ai criteri dati- la focalizzazione sul compito: Definendo il ruolo come analista, l'IA tenderà a evitare risposte casuali o vaghe, e invece cercherà di attenersi a una valutazione coerente e professionale, come farebbe un vero analista.
La richiesta di usare l'intera scala di valori, quando ritenuto opportuno, è stata fatta per evitare che l'IA tendesse a scegliere valori centrali o neutri per sicurezza, in modo da minimizzare il rischio di errori estremi. Invece, le è stato chiesto esplicitamente di considerare tutta la scala, quando appropriato. Inoltre, questa frase funziona come un promemoria interno per il modello: aiuta l'IA a tenere in mente che esiste una scala specifica con significati distinti associati a ciascun valore e che deve essere consapevole dell'uso appropriato di ciascun valore.
Successivamente, è stato creato un ciclo for che, in questo caso, esegue un solo ciclo ($range(1)$), ma funge da struttura per test futuri, nel caso si volesse far eseguire il questionario all'IA un numero elevato di volte. All'interno del ciclo, viene creata una nuova colonna nel DataFrame chiamata "GPT-test-0" e tutti i valori iniziali della colonna sono impostati su 0. In altre parole, viene eseguita un'operazione che aggiunge una nuova colonna al DataFrame e la inizializza con il valore zero (0), questo in particolare serve a preparare il DataFrame per riempirlo successivamente con valori calcolati (il punteggio generato dall'analisi delle frasi) e ad assicurarsi che la colonna esista prima di eseguire ulteriori operazioni, evitando così errori di accesso a colonne non definite. Il successivo ciclo for scorre riga per riga all'interno del DataFrame. Per ogni riga, si estrae il testo

della frase (row['text']). L'indice idx rappresenta il numero della riga. La frase viene stampata per mostrarla all'utente.
Nella sezione successiva viene eseguita la chiamata API. È specificata la versione di ChatGPT da utilizzare (al momento del test la versione 4o risulta essere la più recente). La chiamata all'API include il contesto ($system_instruction$) definito all'inizio ed il messaggio in cui si chiede di utilizzare la scala di valori fornita. È presente, inoltre, il parametro "temperature" che permette di regolare la creatività della risposta del modello. Il modello fornisce una risposta che viene convertita in un numero intero (int) e viene memorizzato nella cella corrispondente del DataFrame. La risposta viene anche stampata.
Se si verifica un errore durante la chiamata all'API, il codice stampa un messaggio di errore. Questo aiuta a diagnosticare problemi in fase di esecuzione del programma.

$time.sleep(1)$: questa pausa serve a garantire che l'intervallo tra una richiesta e l'altra sia sufficiente a rispettare i limiti di frequenza stabiliti dall'API. I limiti di frequenza sono delle regole imposte dai provider delle API per prevenire un sovraccarico del server e garantire un'allocazione equa delle risorse tra tutti gli utenti. Senza la pausa, il ciclo nel codice invierebbe richieste molto rapidamente, probabilmente una dietro l'altra, superando i limiti di frequenza dell'API. Di conseguenza, l'API potrebbe rifiutare le richieste aggiuntive, generare errore o persino penalizzare temporaneamente l'account bloccando ulteriori richieste per un certo periodo. enalizzare temporaneamente l'account bloccando ulteriori richieste per un certo periodo.

print(f"\nCompletato test {i}") stampa, alla fine del ciclo, un messaggio che indica il completamento del test i-esimo, anche questa riga di codice è stata inserita nell'ottica di far rispondere al questionario più volte in modo automatico, qualora sia richiesto.
Infine, il codice tenta di salvare il DataFrame aggiornato in un file Excel. Se il salvataggio riesce, stampa un messaggio di conferma; altrimenti, stampa un messaggio di errore.

CAPITOLO 3

RISULTATI

3.1 Risultati dell'IA: analisi dati forniti da ChatGPT

Il test è stato sottoposto a ChatGPT per un totale di 6 volte, impostando diversi valori di temperatura. due volte la temperatura è stata impostata a 0, tre volte a 0.5 ed una volta a valore 1. I testi a cui ChatGPT ha assegnato un valore sono stati 151, cinque in meno del test condotto su esseri umani. In particolare, i testi erano suddivisi nel seguente modo: 55 appartenenti alla categoria femminile, 49 neutra, 47 maschile. La motivazione per cui hatGPT non ha assegnato un valore cinque testi del dataset va ricercata nella formattazione del Dataset di partenza. Nonostante ciò, è stato ritenuto che l'assenza di cinque testi, partendo da un totale di 156, non invalidasse il risultato ottenuto.

	Category	Expected_Mean	GPT_Mean	Discrepancy_Mean	Discrepancy_Std	Discrepancy_Min	Discrepancy_Max
1	F	-1.5	-1.1090909090909091	0.3909090909090909 3	0.9939209163101397	-0.5	2.5
2	M	1.5	0.723404255319149	-0.776595744680851	0.6821362754188991	-2.5	0.5
3	N	0.0	-0.5510204081632653	-0.5510204081632653	0.9587584091286135	-2.0	2.0

Figura 3.1: Questionario 1, Temperatura 0

A temperatura 0, ad entrambe le esecuzioni del programma sono state fornite 151 risposte, la media complessiva dei valori è stata -0.36 nel primo caso e -0.38 nel secondo, mentre la deviazione standard è stata di 1.17 nel primo e 1.19 nel secondo. Il valore negativo è

dovuto alla presenza in maggior numero di testi rivolti ad un pubblico femminile, in modo compatibile con i risultati del questionario sottoposto ad umani. Facendo riferimento alla fig. 3.1, i risultati dei due test sono stati rappresentati nelle diverse categorie di riferimento ma con un unico valore di media tra le due esecuzioni, dal momento che non vi era quasi alcuna differenza.

Si evince pertanto che risulti essere più rilevante analizzare i risultati in rapporto al valore ipotizzato piuttosto che in senso generale.

1. Temperatura 0,

   Nella **categoria F** (valore atteso: femminile), su un totale di 55 testi:

   - GPT Mean: -1.11: Il valore medio delle valutazioni assegnate da GPT è coerente con le aspettative, le frasi sono tendenzialmente identificate come rivolte ad un pubblico femminile.;
   - Discrepancy Std: 0.99: La deviazione standard di quasi 1 indica una variabilità significativa nelle discrepanze, con alcune frasi valutate molto lontane dal valore atteso.;
   - Discrepancy Max: 2.5: La discrepanza massima di 2.5 indica che ci sono stati casi in cui le valutazioni di GPT si sono discostate notevolmente, classificando alcune frasi più neutrali o addirittura maschili.;

   Nella **categoria N** (valore atteso: neutro), su un totale di 49 testi:

   - GPT Mean: -0.55: GPT tende a classificare le frasi neutre leggermente verso valori più femminili.;
   - Discrepancy Mean: -0.55: La discrepanza media negativa conferma che GPT vede queste frasi come meno neutre e più orientate al pubblico femminile.;
   - Discrepancy Std: 0.96: La deviazione standard mostra una notevole variabilità, suggerendo che GPT a volte classifica frasi neutre con un'ampia gamma di giudizi.
   - Discrepancy Min: -2.0: Alcune frasi sono state valutate come fortemente femminili, molto lontano dall'atteso neutrale. Discrepancy Max: 2.0: D'altro canto, ci sono frasi che GPT ha considerato fortemente maschili, mostrando una significativa fluttuazione nel giudizio.;

È decisamente rilevante notare come, isolando le frasi il cui valore dato da ChatGPT è diverso da quello atteso ed inoltre diverso da 0 (è stato ritenuto meno rilevante che una frase si discosti dal valore atteso ma venga identificata come rivolta ad un pubblico neutro, infatti per brevi porzioni di testo è un dato più trascurabile), siano presenti 23 frasi di cui ben 20 il genere rilevato è stato femminile e solamente 3 maschile. Questo dato evidenzia una sproporzione decisamente inattesa e potrebbe far pensare che dei bias di genere siano presenti nei risultati. I dati suggeriscono che GPT mostra un bias sistematico verso una leggera femminilizzazione delle frasi neutre e una neutralizzazione delle frasi maschili. Questo potrebbe essere dovuto alla natura dei dati su cui il modello è stato addestrato.

	Category	Expected_Mean	GPT_Mean	Discrepancy_Mean	Discrepancy_Std	Discrepancy_Min	Discrepancy_Max
1	F	-1.5	-1.1273	0.3727	1.0193	-0.5	2.5
2	M	1.5	0.6596	-0.8404	0.7879	-3.5	0.5
3	N	0.0	-0.6122	-0.6122	0.975	-2.0	2.0

Figura 3.2: Questionario 1, Temperatura 0.5

2. Temperatura 0.5, **Media delle tre esecuzioni**:
   Nella **categoria F** (valore atteso: femminile), su un totale di 55 testi:

   - Il test condotto a temperatura 0.5 risulta praticamente identico nei risultati rispetto a quello precedentemente eseguito a temperatura 0;

   Nella **categoria N** (valore atteso: neutro), su un totale di 49 testi:

   - GPT Mean: Precedente: -0.55. Attuale: -0.6122. C'è una leggera tendenza a valutare le frasi neutre come più femminili, con un bias leggermente aumentato rispetto al test precedente.;
   - Discrepancy Mean: Precedente: -0.55. Attuale: -0.6122. La discrepanza media più negativa nel test attuale riflette una maggiore tendenza a valutare le frasi neutre come orientate al femminile.;

   Nella **categoria M** (valore atteso: maschile), su un totale di 47 testi:

   - Discrepancy Mean: Precedente: -0.77. Attuale: -0.8404. La discrepanza media negativa è aumentata, suggerendo una tendenza ancora più marcata a sottovalutare la mascolinità.;

- Discrepancy Std: Precedente: 0.68. Attuale: 0.7879. Un leggero aumento nella deviazione standard indica una maggiore variabilità nel giudizio attuale rispetto al precedente.;

Considerazioni generali: si può notare una coerenza e bias persistenti. In entrambi i test, GPT mostra un bias sistematico verso la neutralizzazione delle frasi maschili e la femminilizzazione delle frasi neutre. Questo bias persiste e, in alcuni casi, si intensifica nell'attuale test. è presente un incremento nella variabilità: Un leggero aumento nella deviazione standard per le categorie maschili e neutre suggerisce che il modello potrebbe essere più incerto o meno coerente nelle sue valutazioni rispetto al test precedente. Effettivamente, impostando la temperatura a 0.5 invece che 0, si ottiene un leggero aumento della variabilità nelle risposte ottenute, nello specifico sono accentuati i bias. In sintesi, il test attuale conferma i trend osservati nel test precedente, con alcune aree di peggioramento nella coerenza delle valutazioni, specialmente per le frasi maschili. Rimane fondamentale lavorare su strategie di miglioramento per ridurre il bias e migliorare la precisione delle valutazioni del modello. Anche in questo caso, è decisamente rilevante notare come, isolando le frasi il cui valore dato da ChatGPT è diverso da quello atteso ed inoltre diverso da 0 (è stato ritenuto meno rilevante che una frase si discosti dal valore atteso ma venga identificata come rivolta ad un pubblico neutro, infatti per brevi porzioni di testo è un dato più trascurabile), siano presenti 26 frasi di cui ben 22 il genere rilevato è stato femminile e solamente 4 maschile.

	Category	Expected_Mean	GPT_Mean	Discrepancy_Mean	Discrepancy_Std	Discrepancy_Min	Discrepancy_Max
1	F	-1.5	-1.0545	0.4455	0.9703	-0.5	2.5
2	M	1.5	0.6809	-0.8191	0.8104	-3.5	0.5
3	N	0.0	-0.5102	-0.5102	0.8926	-2.0	2.0

Figura 3.3: Questionario 1, Temperatura 1

1. Temperatura 1

- Categoria **F**: GPT Mean: -1.0545: GPT tende a valutare le frasi meno intensamente femminili rispetto all'atteso, simile ai risultati precedenti, ma con un bias leggermente ridotto. nonostante ciò, la Discrepancy Mean risulta essere

0.4455. Una discrepanza media più alta rispetto alle analisi precedenti, suggerendo una deviazione leggermente più marcata dalle aspettative. Questo dato è compatibile con l'aumento ad 1 nel valore di temperatura.

- Categoria **N**: In questo caso, il test è risultato leggermente più preciso nell'indentificazione di testi rivolti ad un pubblico neutro.
- Categoria **M**: GPT Mean: 0.6809. GPT sottovaluta la mascolinità delle frasi, un trend costante rispetto ai test precedenti, ma con una valutazione media leggermente superiore. La discrepanza media è simile al test precedente, indicando una persistente tendenza a valutare le frasi maschili come meno maschili del previsto.

Anche in questo caso, coerentemente con i test precedenti, i risultati mostrano che ci sono un numero significativo di frasi con punteggi che non corrispondono ai valori attesi. In questo caso, isolando le frasi il cui valore dato da ChatGPT è diverso da quello atteso ed inoltre diverso da 0 (è stato ritenuto meno rilevante che una frase si discosti dal valore atteso ma venga identificata come rivolta ad un pubblico neutro, infatti per brevi porzioni di testo è un dato più trascurabile), siano presenti 22 frasi di cui ben 19 il genere rilevato è stato femminile e solamente 3 maschile.

3.2 Confronto tra risposte umane e IA

categoria	n° sezioni	n° medio risp./domanda	score medio	deviazione std	min	max
neutra	52	6.29	-0.1774	0.4847	-1.5	+0.8
femminile	55	6.62	-0.7001	0.6703	-1.857	+1
maschile	49	6.02	+0.5109	0.8084	-1.6	2

Figura 3.4: Questionario 1, esseri umani

categoria	n° sezioni	score medio	deviazione std	min	max
neutra	55	-0,6122	0,975	-1,7	0,9
femminile	49	-1,1273	1,0193	-1,8	1,2
maschile	47	0,6596	0,7879	-1,4	2

Figura 3.5: Questionario 1, Temperatura 0.5

Confrontando i valori ricavati sottoponendo il questionario agli esseri umani, visibili in 3.6, con quelli ottenuti da IA, possiamo fare alcune interessanti osservazioni. Per quanto riguarda i testi rivolti ad un pubblico neutro, il valore medio di GPT suggerisce un bias più marcato verso valori femminili rispetto agli umani, che sono leggermente più neutrali. La deviazione standard inoltre mostra una maggiore variabilità, suggerendo un'incertezza più elevata dell'IA nella classificazione delle frasi neutre. Min/Max: GPT ha un range più ampio nelle valutazioni, con casi estremi che raggiungono sia valutazioni femminili che maschili. Per quanto riguarda le frasi rivolte ad un pubblico femminile, considerando il valore medio si nota che GPT tende a valutare le frasi più vicine all'atteso rispetto agli umani, che mostrano una tendenza ad maggiore ad assegnare valori neutri. Tuttavia, GPT presenta una maggiore variabilità nelle valutazioni rispetto agli umani, suggerendo incertezza o incoerenza. Questo ci porta a dire che gli esseri umani hanno un range di valutazione più contenuto, mentre GPT mostra casi estremi con valutazioni molto maschili. Infine, confrontando i valori dei testi di categoria maschile, si evince che GPT attribuisca mediamente un valore più spostato verso valori positivi, quindi effettivamente maschili, ma con valori Min/Max che mostrano una discrepanza minima decisamente più estrema rispetto agli umani, indicando una maggiore tendenza a valutare alcune frasi come molto femminili. in generale, GPT tende a spostare le frasi maschili e neutre verso valori meno intensi rispetto agli umani. GPT mostra inoltre una maggiore variabilità nelle valutazioni, suggerendo che il modello ha difficoltà a stabilire un criterio coerente per giudicare le frasi di genere. é chiaro che le risposte ottenute da IA tendano, a tutti i valori di temperatura testati, ad attribuire in modo decisamente inaspettato molte più frasi come rivolte ad un pubblico femminile. non si può dire in che misura questo esito sia da attribuire ad un bias presente nell'IA e quanto al Dataset di testi utilizzato. Pertanto si è pensato di costruire un nuovo dataset aggiornato, con una metodologia di selezione delle frasi leggermente diversa, ed eseguire nuovamente il questionario. Nonostante una marcata tendenza di ChatGPT nel femminilizzare frasi che gli esseri umani ritengono rivolte ad un pubblico più neutro, non si può dire lo stesso per quanto riguarda le frasi ipotizzate per un pubblico maschile. Ciò è chiaro se si osserva il grafico in fig. 3.6, che rappresenta le risposte di Chatgpt (sull'asse delle ascisse x) in relazione al punteggio ottenuto da umani (sull'asse delle ordinate y). Si nota infatti che le frasi ipotizzate come "maschili" hanno ottenuto un punteggio molto coerente, con valori che si avvicinano alla retta e quasi mai scendono nel quadrante negativo, che rappresenta le frasi viste come "femminili". Al contrario, andando a studiare le frasi del dataset che dovrebbero rivolgersi ad un pubblico

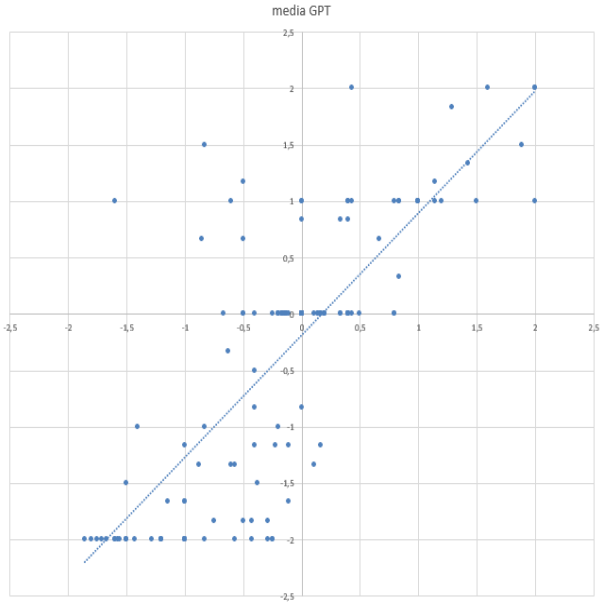

Figura 3.6: Questionario 1, confronto GPT-umani

femminile, si notano delle differenze interessanti. Da un lato, come già evidenziato in precedenza, c'è una maggior tendenza di GPT ad identificare frasi come fortemente rivolte ad un pubblico femminile, sbilanciandosi quindi in modo maggiore rispetto agli umani. Dall'altro lato, sono presenti alcuni valori in forte disaccordo, che pertanto vedremo nello specifico. La frase, presa da un articolo per uomo,: "Declinata in versione Eau de Parfum nel flacone della Skyline Collection, Kashan Oudh di Locherber Milano è la fragranza misteriosa e mistica con cui risvegliare in lui suggestioni lontane grazie alle note dolci, speziate e legnose." è stata valutata dagli umani come particolarmente rivolta ad ub pubblico femminile (media -1.6), tuttavia ha ricevuto punteggio +1 in tutte le esecuzioni del programma con GPT. questo risultato potrebbe spiegarsi con la presenza del pronome "lui", la possibilità che GPT abbia identificato il profumo come "per uomo", mentre gli umani si siano affidati al contesto. Un altro esempio, ancor più significativo, lo si trova in un'altra frase rivolta a pubblico maschile: "Nel frattempo, secondo questa logica distorta, sono le femmine a potere scegliere. Una visione universale ancora più tetra e pericolosa della red pill è quella derivante dalla black pill, la pillola nera". il termine "red pill" deriva dal concetto di "pillola rossa" reso popolare dal film Matrix (1999). Nel film, prendere la pillola rossa significa accettare una realtà scomoda e sconvolgente, in contrapposizione alla pillola blu, che rappresenta la scelta di rimanere in un'illusione confortevole. Nella società moderna ed in particolare nei social network, i redpillati (come si autodefiniscono) discutono temi legati alle dinamiche di genere, criticando le aspettative sociali verso uomini e donne e promuovendo visioni che sostengono che gli uomini siano

svantaggiati o manipolati dalle donne o dalla società moderna, spesso sfociando in vero e proprio odio misogino. La frase è stata valutata come rivolta ad un pubblico femminile dagli umani (media -0.83), al contrario come fortemente rivolta ad un pubblico maschile da GPT (media +1.5).

3.3 Nuovo dataset

I criteri di selezione per la scelta degli articoli sono stati: tipologia di fonti varia, quindi sono stati inclusi articoli accademici, giornalistici e riviste. La lingua degli articoli è l'italiano. Il periodo di pubblicazione rientra negli ultimi 5 anni (questo criterio di inclusione è volto a creare un database di testi il cui linguaggio sia quanto più attuale possibile). Gli articoli sono disponibili gratuitamente online.

1. Raccolta documenti testuali, alcuni esempi

   Articoli rivolti ad un pubblico maschile:

   - **GQ Italia**: acronimo di "Gentleman's Quarterly" magazine dedicato alla moda maschile, sport e tendenze.
   - **Uomo — Vogue Italia**: una delle principali pubblicazioni nel settore della moda maschile.

   Articoli rivolti ad un pubblico neutro:

   - **DigitCult**: rivista accademica che raccoglie pubblicazioni sul cambiamento sociale, la cultura digitale e l'innovazione tecnologica.
   - **Gravità Zero**: Blog di divulgazione scientifica.
   - **Il Bo Live**: la rivista curata dall'università di Padova (www.ilbolive.unipd.it).

   Articoli rivolti ad un pubblico femminile:

   - **Donna Moderna**: rivista che pubblica articoli su salute, benessere, moda e lifestyle per donne.
   - **Marie Claire**: rivista che tratta un mix di moda, bellezza, cultura e tematiche sociali e politiche.
   - **Io Donna**: sezione de "Il Corriere Della Sera" dedicata ad un pubblico femminile.

Lunghezza ottimale: 30-80 parole per porzione di testo. Motivazione: Questa lunghezza è abbastanza breve da essere letta e compresa rapidamente. Un testo più breve potrebbe non fornire abbastanza informazioni per una risposta accurata, mentre un testo troppo lungo potrebbe affaticare i partecipanti e ridurre la qualità delle risposte. Metodo di selezione: è stato chiesto a chatgpt di selezionare in modo casuale delle porzioni di testo di 30-110 parole, di senso compiuto, dagli articoli interi. Da queste sono state poi estratte porzioni di testo più contenute (30-80). Questo per evitare un mio possibile bias nella scelta delle porzioni di testo (che io potessi scegliere le frasi pensando fossero indirizzate al sesso coerente con quello di indirizzo della categoria). La scelta di tagliare testo da uno più lungo di partenza è dovuta al fatto che non sempre chatgpt era in grado di fornire frasi adeguatamente comprensibili (es. frasi che iniziano con "proprio per questo possiamo dire che.."). L'elenco completo dei testi è disponibile[13].

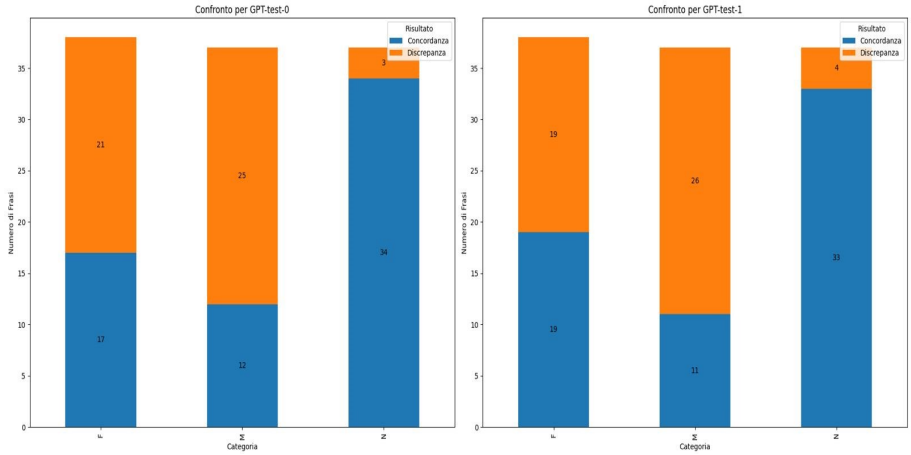

Figura 3.7: Esiti del secondo questionario

Il questionario è stato sottoposto a ChatGPT un totale di quattro volte, la temperatura è stata impostata a 0 e 0.5, anche in questo caso non si evidenziano differenze degne di nota per i motivi precedentemente descritti. Facendo riferimento alla 3.5, che riporta dati relativi a due esecuzioni del programma, poichè in generale i risultati non hanno presentato significative differenze, si può notare come i testi il cui genere ipotizzato è neutro vengano rilevati con precisione sensibilmente più accurata, quasi totale.

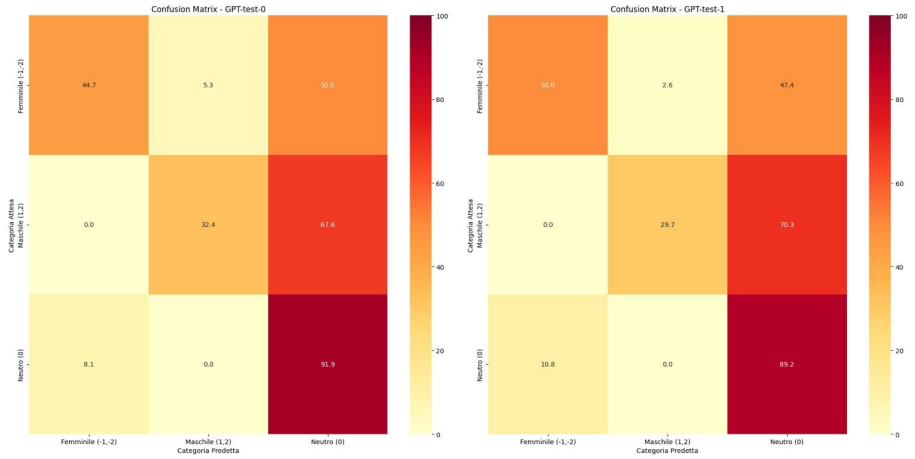

Figura 3.8: Matrice di confusione

La matrice di confusione in Fig. 3.8 è uno strumento utilizzato per valutare le performance di un modello di classificazione. È una tabella che permette di visualizzare le prestazioni del modello confrontando i valori predetti con quelli effettivi (o veri) delle classi. I valori confermano nuovamente che la percentuale di testi neutri rilevati comne tali è molto elevata. Si può notare inoltre come la maggior parte dei testi che hanno ottenuto una catalogazione differente da quella aspettata rientrino nel genere neutro, un risultato in linea con le aspettative poichè le frasi sono state selezionate per non dare un eccessivo contesto dell'articolo originale.

	Category	Expected_Mean	GPT_Mean	Discrepancy_Mean	Discrepancy_Std	Discrepancy_Min	Discrepancy_Max
1	F	-1.5	-0.6842	0.8158	1.0425	-0.5	3.5
2	M	1.5	0.4054	-1.0946	0.6438	-1.5	0.5
3	N	0.0	-0.1622	-0.1622	0.5534	-2.0	0.0

Figura 3.9: Dati numerici

Infine, presentimo i dati di media e deviazione standard dei risultati. confrontando l'expected mean: -1.5 (valore atteso medio per frasi rivolte a un pubblico femminile) con GPT Mean: -0.75 (valore medio assegnato dal modello GPT) si evince che il modello tende a

classificare frasi attese come "femminili" più vicino al neutro o in alcuni casi verso il maschile. Il dato di discrepancy Min/Max: Da -0.5 a +2.5 indica che le valutazioni possono occasionalmente essere molto lontane dall'atteso. i risultati per la categoria maschile indicano meno variabilità rispetto alla categoria femminile, ma ancora presente, come si nota da discrepancy Std: 0.60 e discrepancy Min/Max: Da -1.5 a +0.5. in generale si evince una minore precisione nelle categorie estreme: Le frasi femminili e maschili mostrano discrepanze più marcate rispetto a quelle neutre, suggerendo che il modello ha difficoltà a catturare il genere estremo nel linguaggio. Rispetto all'esecuzione del test con il dataset precedente, si può notare che, nella categoria maschile la valutazione media si è spostata verso valori neutri. La variabilità è minore, suggerendo valutazioni più coerenti. I valori estremi minimi sono migliorati rispetto al test precedente, con meno valutazioni estremamente femminili. Per quanto riguarda la categoria neutra, si vede chiaramente che il test condotto sul nuovo dataset ha migliorato la valutazione media verso valori effettivamente più neutri. con una discrepanza media che è notevolmente ridotta, avvicinandosi ai valori attesi. Seguono lo stesso trend i risultati per la categoria femminile, infatti Il test mostra una maggiore tendenza a spostarsi verso valori neutri, con una maggiore variabilità nelle valutazioni. In generale, il test presenta una maggiore coerenza nelle valutazioni neutre ed una maggior tendenza generale a valutrare i testi come neutri.

CONCLUSIONI E FUTURI SVILUPPI

L'analisi evidenzia che, sebbene GPT mostri alcune capacità nel riconoscere e classificare il genere delle frasi, ci sono aree significative di miglioramento, in particolare nella coerenza delle valutazioni e nella gestione dei bias. Investire in un addestramento più equilibrato e in tecniche di miglioramento del contesto potrebbe portare a un modello più robusto e accurato nel trattamento delle frasi di genere. È necessario un affinamento del modello per migliorare la sua capacità di riconoscere e rispettare l'intensità di genere delle frasi. Questo potrebbe includere un ribilanciamento del Dataset: Garantire che il dataset di addestramento contenga una rappresentanza equa e diversificata di frasi connotate in termini di genere. Inoltre sarebbe opportuno un miglioramento del Contesto: Addestrare il modello con un contesto più ricco che permetta di cogliere meglio le sfumature di genere. Creare metriche specifiche per misurare la precisione e la coerenza delle valutazioni di genere potrebbe facilitare il monitoraggio dei miglioramenti e l'identificazione delle aree problematiche. Un aspetto interessante a tal proposito sono i metodi di ensemble, ovvero tecniche di apprendimento automatico che combinano le predizioni di più modelli per migliorare le prestazioni complessive rispetto a quelle ottenute da un singolo modello. L'idea alla base è che diversi modelli possano catturare diverse caratteristiche dei dati e, combinando le loro predizioni, si ottiene un risultato più robusto e accurato. I modelli di ensemble sono più complessi e richiedono più risorse computazionali, ma restano un'importante sfida per il futuro. Un interessante sviluppo del presente test potrebbe essere quello di sottoporre il questionario ad altre IA trainate in modi differenti oppure che presentino caratteristiche distinte da ChatGPT, ad esempio Llama di Meta, che è una AI open source. In conclusione, al momento risulta difficile rispondere alla domanda se l'AI possa essere un valido sostituo all'essere umano in interazioni come quello di

creare un Dataset privo di bias, i risultati ottenuti nel test mostrano due tendenze opposte, nel femminilizzare eccessivamente un numero non indifferente di frasi e, al contrario, nell'identificare correttamente testi che ottengono una valutazione umana in forte disaccordo con le aspettative. L'aumento dei dati ed il miglioramento delle tecniche di training rappresenteranno un aspetto fondamentale per rispondere al quesito.

BIBLIOGRAFIA

[1] A. M. Turing (1950) Computing Machinery and Intelligence. Mind 49: 433-460. (1950) https://courses.cs.umbc.edu/471/papers/turing.pdf

[2] Stuart Russell, Of Myths and Moonshine, contributo alla conversazione The Myth of AI. Disponibile all'indirizzo: https://www.edge.org/conversation/the-myth-of-ai#26015, 26/01/2022.

[3] Homo Heuristicus: Why Biased Minds Make Better Inferences Gerd Gigerenzer, Henry Brighton First published: 30 January 2009 https://doi.org/10.1111/j.1756-8765.2008.01006.x

[4] Eichler, M., Reisman, A. L., & Borins, E. M. (1992). Gender Bias in Medical Research. *Women & Therapy*, 12(4), 61–70. https://doi.org/10.1300/J015v12n04_06

[5] Savoldi, B., M. Gaido, L. Bentivogli, M. Negri, M. Turchi (2021), 'Gender Bias in Machine Translation', Transactions of the Association for Computational Linguistics, 9, pp. 845-874. https://aclanthology.org/2021.tacl-1.51/.

[6] Ferrara, Emilio, Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models. Available at SSRN: https://ssrn.com/abstract=4627814 or http://dx.doi.org/10.2139/ssrn.4627814.

[7] arXiv:2410.20739 [cs.CL] disponibile all'indirizzo https://doi.org/10.48550/arXiv.2410.20739

[8] Vincenzo Ambriola, Agi (intelligenza artificiale generale): da dove arriva e dove va. Pubblicato il 28 feb 2024, https://www.agendadigitale.eu/cultura-digitale/da-turing-ai- modelli-generativi-percorsi-e-sfide-dellintelligenza-artificiale-generale/.

[9] Artuficial analysis.ai https://artificialanalysis.ai/models/gpt-4o-2024-08-06.

[10] Scagnet Martino, Sviluppo e validazione di un dataset in italiano per l'analisi di stereotipi di genere nei documenti testuali https://thesis.unipd.it/bitstream/20.500.12608/48854/1/Scagnet_Martino.pdf.

[11] Silvana Badaloni, Antonio Rodà1 and Martino Scagnet AnItalian dataset for the analysis of gender stereotypes in textual documents https://ceur-ws.org/Vol-3615/short1.pdf.

[12] Italian Gender Bias Dataset https://zenodo.org/records/10027951.

[13] Codice Python https://colab.research.google.com/drive/1ertxNp5YOOgYzCOFReeRrTY7vJWSDhQ4.

[14] Dataset2 https://docs.google.com/spreadsheets/d/1Ei3Mr8hTK8k6crrvV_vpGL5Kj7MJa36u/edit?gid=266423866#gid=266423866.

RINGRAZIAMENTI

Beh, papone.. grazie per il supporto, per l'amore e i consigli che hai saputo darmi, sempre. nei momenti peggiori avevi sempre la frase giusta che mi faceva sbloccare e andare avanti. mi hai dato la tua forza anche quando io non ne avevo (spesso haha). grazie, non so quantificare il bene che ti voglio. E poi ci sono i miei amici. dalla sbg alla stk, siete tanti bastardoni che mi avete fatto vivere anni veramente divertenti e indimenticabili. Grazie giovi senza i tuoi svasi non sarei neanche arrivato qui probably. gtex, daddy berzi vio mike violetta maga riki simo amy gianni siete tanti e non sto a fare l'elenco completo, voi sapete che ne abbiamo passate di storie, grazie a quelli che vedo ancora e chi meno (gilda riki giada ecccccccc). e infine grazie a te vitty, che non mi conosci neanche e mi conosci più di tutti, che mi hai fatto capire tante cose e mi hai lasciato con altrettanti punti interrogativi.