**Bias di genere e intelligenza artificiale: una mancanza di competenze?**

Mara Floris
*Università Vita-Salute San Raffaele*

Nel novembre del 2017, lo scrittore americano Alex Shams ha creato un certo scalpore su diversi social media che hanno ripreso un *tweet* nel quale scriveva: «Il turco è una lingua neutra dal punto di vista del genere. Non ci sono 'lui' o 'lei': tutto è semplicemente 'o'. Ma guardate cosa succede quando Google traduce in inglese»[1]. Il testo era seguito da uno *screenshot*, nel quale alcune brevi frasi contenenti il pronome turco neutro "o" sono tradotte dal turco all'inglese, ottenendo il risultato in figura 1:

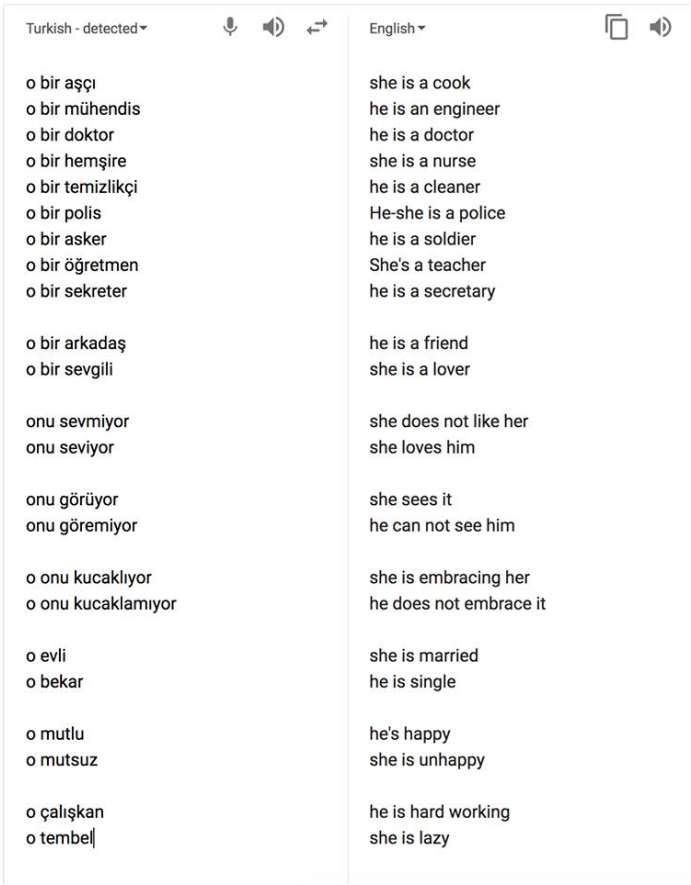


*Figura 1. Screenshot del tweet di Alex Shams del 28 novembre 2017* (fonte: https://twitter.com/alexshams_/status/935291317252493312

I pronomi personali vengono tradotti al maschile o al femminile a seconda dello stereotipo di genere relativo alla professione indicata: "ingegneri" e "dottori" sono uomini, "infermiere" e "segretarie" donne. Lo stesso succede quando ai pronomi personali vengono accostati sentimenti e atteggiamenti: le traduzioni rinforzano lo stereotipo di donna fragile ed emotiva (Prates *et al.* 2020).

In letteratura, la presenza di un *bias* di genere nelle traduzioni automatiche e in altri sistemi di *Natural Language Processing* (NLP) è ampiamente attestata (Chen *et al.* 2021; Costa-jussà 2019;

[1] «Turkish is a gender neutral language. There is no 'he' or 'she' — everything is just 'o'. But look what happens when Google translates to English» (traduzione mia).

Sun *et al.* 2019) Il problema è duplice: manca una formazione specifica sui *bias* di genere e mancano anche competenze linguistiche specifiche sulle lingue diverse dall'inglese.
Dopo le numerose segnalazioni, Google ha provveduto a correggere il tiro. Attualmente, se provate a tradurre dal turco "*o bir doktor*", otterrete "lei è un dottore". Molto meglio, anche se il termine corretto in italiano è "dottoressa", un errore che avrebbe potuto essere evitato facilmente da una persona con le competenze adeguate dal punto di vista linguistico e con un'appropriata formazione sull'esistenza di un *bias* di genere.
Un'area in cui i *bias* di genere sono particolarmente diffusi è quella dell'elaborazione del linguaggio naturale (NLP) e delle traduzioni (cfr. Luccili *et al.* 2020). Per *bias* di genere si intende il favoritismo, o la discriminazione sistematica, nei confronti di un particolare genere, con conseguente disparità di trattamento e di opportunità. Sebbene nelle dichiarazioni degli intenti dei produttori l'intelligenza artificiale (IA) miri a essere obiettiva e imparziale, spesso rispecchia i *bias* presenti nei dati su cui viene addestrata. I dati utilizzati per addestrare i modelli di IA possono contenere preconcetti, ereditati dalla società in cui è stata progettata e che portano a risultati distorti.
Oltre al summenzionato caso della traduzione, un altro caso in cui è presente un *bias* di genere sono i sistemi di *sentiment analysis*. L'analisi dei sentimenti di un testo effettuata da un sistema di NLP potrebbe associare erroneamente parole o espressioni tipicamente utilizzate da donne o uomini con sentimenti positivi o negativi, creando un'interpretazione distorta e rinforzano stereotipi di genere.
Altri esempi della presenza di tale problema nel NLP sono i sistemi di *screening* dei *curricula*. Se un sistema di selezione automatizzata dei *curricula* addestrato su dati storici mostra una preferenza per determinati termini o esperienze tipicamente associate a un genere specifico, tale preferenza potrebbe scoraggiare le candidate femminili o maschili a seconda del *bias* incorporato.
Per affrontare in modo completo tutte le possibili conseguenze di tali discriminazioni nel mondo delle tecnologie di NLP, è necessario compiere diversi passi importanti, che includono lo sviluppo di nuove professioni, l'acquisizione di competenze specialistiche da parte dei linguisti e uno sforzo congiunto per interrompere il processo attraverso il quale i *bias* di genere si riflettono nelle tecnologie che produciamo. In tal senso, un aspetto cruciale riguarda la creazione di nuovi profili professionali con una formazione linguistica, in particolare sulle lingue cui è presente il genere grammaticale, e con un'adeguata preparazione sui *bias* di genere e sulle manifestazioni linguistiche di questo fenomeno. Tradizionalmente, lo sviluppo dell'NLP ha fatto affidamento principalmente sulle competenze di informatici e ingegneri, che potrebbero non possedere una comprensione approfondita delle sfumature e della mancanza d'inclusione linguistica. Incorporando specialisti del linguaggio con una formazione adeguata in lingue diverse dall'inglese, possiamo garantire una prospettiva più completa e culturalmente diversificata nel processo di sviluppo. Tali figure possono contribuire con la loro esperienza all'analisi delle strutture linguistiche, con un occhio di riguardo per il problema del genere, all'identificazione di discriminazioni potenziali e alla proposta di strategie per attenuarle o eliminarle.

Inoltre, i linguisti con una formazione specifica in sociolinguistica, analisi del discorso e studi di genere, permetterebbero una comprensione maggiore di come i *bias* di genere si manifestano nel linguaggio. Sfruttando la propria esperienza, possono contribuire attivamente allo sviluppo di algoritmi e modelli più sensibili alle diverse espressioni e identità di genere.
Bloccare il processo attraverso il quale i *bias* di genere si riflettono nelle tecnologie NLP è un obiettivo fondamentale. Ciò comporta la loro identificazione e correzione in varie fasi del processo di sviluppo. Stabilendo metodologie di valutazione rigorose, possiamo valutare sistematicamente i modelli di NLP alla ricerca delle discriminazioni di genere e perfezionarli

iterativamente per garantire equità e inclusione. Inoltre, è essenziale creare insiemi di dati di formazione diversificati e rappresentativi. Ciò può essere ottenuto incorporando le prospettive delle comunità emarginate e consultando individui con diverse identità di genere durante i processi di raccolta e annotazione dei dati. Lavorando attivamente per ridurre i pregiudizi di genere nei sistemi NLP, possiamo promuovere tecnologie più eque e giuste che contribuiscono a una società più inclusiva.

**Bibliografia**

Chen Y., Mahoney C., Grasso I., Wali E., Matthews A., Middleton T., Matthews J. (2021), "Gender bias and under-representation in natural language processing across human languages", in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 24-34.

Costa-jussà Marta R. (2019), "An analysis of gender bias studies in natural language processing", *Nature Machine Intelligence*, 1(11), 495-496.

Luccioli Alessandra, Ester Dolei, Chiara Xausa (2020), "Investigating Gender Bias in Machine Translation. A Case Study between English and Italian", in Adriano Ferraresi, Roberta Pederzoli, Sofia Cavalcanti, Randy Scansani (a cura di), *MediAzioni* 29: B29-B49, http://www.mediazioni.sitlec.unibo.it

Prates M. O., Avelar P. H., Lamb L. C. (2020), "Assessing gender bias in machine translation: a case study with google translate". *Neural Computing and Applications*, 32, 6363-6381.

Sun T., Gaut, A., Tang S., Huang Y., ElSherief M., Zhao J., Wang W. Y. (2019), "Mitigating gender bias in natural language processing: Literature review", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://aclanthology.org/P19-1159/