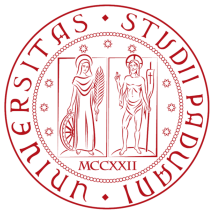

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in

Ingegneria Informatica

Pregiudizi di genere nei Word Embeddings: verso un'analisi del gender score nei documenti testuali.

Relatore

Prof. Antonio Rodà

Correlatrice

Prof.ssa Silvana Badaloni

Laureando

Luca Friso

1236455

Anno Accademico 2021-2022

12 Dicembre 2022

C'è vero progresso solo quando
i vantaggi di una nuova tecnologia diventano per tutti.
— Henry Ford

Indice

Elenco delle figure

Elenco delle tabelle

Abstract

I pregiudizi di genere sono, ad oggi, una delle tematiche più studiate e trattate in vari campi di ricerca, dalla psicologia all'informatica, dalle scienze umanistiche alla matematica.
Questo studio cerca di fornire una panoramica sulle principali cause di *bias* di genere nei documenti di testo scritti in lingua italiana. Attraverso l'utilizzo di tecniche come i *Word Embeddings*, gli *n-grams* e i *Sentence Embeddings* è possibile determinare il *gender level* di un testo e di conseguenza definire una possibile strategia per ridurre la presenza di stereotipi di genere.
Dopo una prima parte di studio delle principali tecniche e metodologie per la determinazione dei *bias* nei documenti di testo, viene presentata una metodologia da seguire per confrontare le tre tecniche indicate in precedenza e determinare se una di queste possa catturare in maniera più accurata il livello *gender* del documento analizzato.
Dai risultati ottenuti sui trentotto documenti analizzati, suddivisi equamente in tre categorie: *gender female*, *gender male* e *gender neutral*, emerge che i *Word Embeddings* sono lo strumento più performante dei tre analizzati, mappando correttamente ventisette documenti su trentatré. Seguono n-grams e *Sentence Embeddings*, che mappano correttamente rispettivamente sedici e quindici documenti.

Introduzione

Negli ultimi anni le applicazioni basate sull'analisi semantica del testo hanno visto un notevole sviluppo, supportate dall'implementazione di sistemi di Intelligenza Artificiale sempre più avanzati e performanti. Questa costante ricerca nel campo del Natural Language Processing (NLP) è stata, e viene tutt'ora, utilizzata per cercare di rendere più performanti *task* come il miglioramento della ricerca web, il perfezionamento della comprensione del linguaggio parlato da parte degli assistenti vocali, il *parsing* dei Curriculum Vitae o la ricerca di commenti denigratori sui *social network*.

Uno strumento largamente utilizzato in queste ricerche sono i Word Embeddings, a cui faremo riferimento con l'abbreviazione WEs, ovvero un *framework* che permette la rappresentazione di parole come vettori. Questa tecnica ha permesso di dimostrare che spesso le parole portano con loro stereotipi di genere che i WEs a volte amplificano. Diversi studi dimostrano la presenza di stereotipi di genere nei vettori rappresentanti parole (Bolukbasi, Chang, Zou, Saligrama, Kalai, 2016; Mikolov, Chen, Corrado, Dean, 2013; Schmidt, 2015; Mikolov, Yih, Zweig, 2013; Nissim, Noord, van der Goot, 2019; Bolukbasi, Chang, Zou, Saligrama, Kalai, 2016; Dev, Phillips, 2019; Garg, Schiebinger, Jurafsky, Zou, 2018). Proprio per questo gli studi effettuati sui WEs cercano di mitigare il più possibile eventuali *bias* e stereotipi di genere presenti nel linguaggio scritto e parlato.

Per definire il *gender bias* i WEs utilizzano la *cosine similarity* tra il vettore relativo alla parola analizzata e la direzione che identifica il genere. Maggiore è la distanza fra il vettore e la direzione, maggiore sarà il *bias* che la parola porta con sé.

Questo studio cerca di proporre un metodo per definire il *gender level* di un documento testuale in lingua italiana attraverso l'utilizzo dei WEs. Dopo un primo passaggio nel quale viene processato il documento eliminando quanto più "rumore" possibile, mantenendo solamente le parole che danno un contributo significativo, si passa al calcolo del *gender level* del documento mediando lo *score* ottenuto dalle varie parole rispetto ad una direzione principale. Nonostante ci siano diversi studi che cercano di riprodurre quanto fatto per i WEs di altre lingue utilizzando i WEs italiani, la letteratura italiana sul tema non è così ampia (Biasion, Fabris, Silvello, Susto, 2020). È stato preso come documento di riferimento per lo studio preliminare e la determinazione della direzione di genere da considerare come principale *Gender Bias in Italian Word Embeddings* (2020) di Biasion, Fabris, Silvello, Susto.

All'interno del capitolo 1 si è svolta una ricerca bibliografica e sitografica sulla tematica dei *bias* di genere e gli stereotipi. Dopo una prima analisi sulle definizioni di questi termini, si è cercato di identificare i principali *bias* di genere presenti nell'intelligenza artificiale, analizzando in particolare i *bias* nei *dataset* di addestramento utilizzati per allenare le reti neurali.
Sono state riportate e trattate alcune riflessioni di docenti che sono intervenuti all'interno del corso "Saperi di Genere ed Etica nell'Intelligenza Artificiale" tenuto dalla professoressa Badaloni e dal professor Rodà, afferente alla laurea in Ingegneria Informatica, ma aperto anche a studentesse e studenti appartenenti ad altri corsi di studio e ambiti formativi.
Sono stati riportati diversi casi di studio di strumenti utilizzati che presentavano stereotipi e *bias* di genere come OPTUM, COMPAS e il caso dei Curriculum Vitae di Linkedin.
L'ultimo argomento tratto nel capitolo 2 costituisce un breve approfondimento dei *bias* nei documenti testuali, con alcuni richiami alla letteratura degli studi in questo ambito e alle tecniche maggiormente utilizzate.

Nel capitolo 2 dopo aver introdotto la macro area del Natural Language Proces-

sing, con i relativi campi di utilizzo e i principali task che NLP cerca di automatizzare, è stato definito il modello dei Word Embeddings, tema centrale di questo elaborato, come strumento per la definizione del *gender level* di un documento di testo. Successivamente sono stati analizzati i principali modelli presenti in letteratura per la costruzione dei Wes, ovvero Word2Vec e GloVe, dei quali si sono riportati i concetti di base e le principali caratteristiche.

All'interno del capitolo 3 sono stati presentati i due paper che hanno guidato questo lavoro, ovvero *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (2013) di Bolukbasi, Chang, Zou, Saligrama, Kalai e *Gender Bias in Italian Word Embeddings* (2020) di Biasion, Fabris, Silvello, Susto. Si è cercato di fornire un'idea generale su quanto è stato trovato e testato dagli autori e di spiegare le parti che sono poi state riutilizzate per lo studio di questa tesi.

Il capitolo 4 rappresenta lo snodo centrale di tutto lo studio. Qui vengono riportate le tre analisi principali fatte per determinare il *gender level* di un documento testuale: WEs, N-grams e Sentence Embeddings. Nel dettaglio, sono stati elencati i passaggi effettuati per ognuna delle tre analisi, dalla scelta del modello alla dimensionalità dei vettori considerati, e altre caratteristiche.

Nel capitolo 5 sono stati riportati alcuni dei risultati ottenuti e un confronto tra di essi al fine di determinare l'accuratezza di quanto restituito dalle varie analisi.

Infine, vengono riportate le conclusioni e alcune considerazioni su quanto fatto, sullo stato attuale dei WEs per la lingua italiana e su quali potrebbero essere i futuri *upgrade* in questa area di studio.

Capitolo 1

Bias di genere e stereotipi

I *gender bias* e gli *stereotipi* possono sembrare due macro-tematiche scollegate tra loro, entrambe rappresentanti problemi della società contemporanea, ma senza una forte correlazione. In realtà i *gender bias*, sono generati da stereotipi legati alle differenze di genere.
La parola “stereotipo” deriva da due termini greci, “stereos” e “typos”, letteralmente “immagine rigida”. La psicologia definisce lo stereotipo come una "rappresentazione mentale che è condivisa da un gruppo sociale e si riferisce ad un altro gruppo sociale a cui vengono attribuite delle caratteristiche che non rispecchiano la realtà, ma solo delle approssimative generalizzazioni"[1].

Alcuni degli stereotipi più frequenti che ritroviamo nella società moderna sono ad esempio che il genere femminile è più adatto all’accudimento rispetto a quello maschile, che le donne sono meno focalizzate sul lavoro dopo la maternità, che le donne sono più predisposte ad affrontare tematiche umanistiche, che il gioco del calcio è uno sport per soli uomini, che l’uomo deve provvedere al fabbisogno della famiglia, che gli uomini hanno un miglior senso dell’orientamento.
Questi *gender bias* si riflettono in vari ambiti della vita, dal lavoro alle relazioni, dalla sfera sociale a quella familiare, causando spesso notevoli problemi a chi è vittima di queste discriminazioni.
Uno dei maggiori problemi che è stato rilevato in diversi studi effettuati negli ultimi

[1] https://www.noemahr.com/gender-bias-cosa-sono/

anni, riportato da agenzie come l'Istat e Almalaurea, ed ampiamente trattato nel corso "Saperi di Genere ed Etica nell'Intelligenza Artificiale", è il numero ancora molto basso di donne che occupano posizioni di vertice o di rilevante responsabilità all'interno delle aziende (Kollmayer, Schober, Spiel, 2018). Dati dell'Istat confermano che l'Italia si trova all'ultimo posto di questa classifica con appena il 48,9% di donne in posizioni rilevanti rispetto al 62,4% della Grecia che occupa il penultimo posto.

La professoressa Caterina Suitner in una delle sue lezioni di Saperi di Genere ed Etica nell'Intelligenza Artificiale ha evidenziato temi quali la segregazione lavorativa, ossia l'esclusione da alcune categorie occupazionali effettuata in base al genere del lavoratore o della lavoratrice, e la visione stereotipata dei ruoli lavorativi, per la quale l'uomo deve avere un alto status sociale, un alto livello di indipendenza e competenza, un basso livello di calore e gentilezza, mentre la donna deve avere un basso status sociale, un basso livello di competenza e indipendenza e un alto livello di calore e gentilezza.

Altro tema cruciale è quello del *soffitto di vetro* ovvero quella "barriera invisibile che impedisce alle donne (e ad altre minoranze) di ottenere posizioni di potere" (Eagly, Carli, 2003, p.807–834). Ciò porta alla *paura del contraccolpo*, ovvero la paura di essere "puniti" per rompere i soffitti di vetro che ci tengono "ai nostri posti", con un conseguente aumento del conformismo e la tendenza a nascondere la devianza. Come accennato in precedenza, questa discriminazione si riversa principalmente sulle donne che, anche se innegabilmente competenti, vengono sminuite, definite fredde e poco apprezzate come persone.

Gli studi riguardanti gli stereotipi, non solamente quelli di genere, hanno inizio già negli anni Trenta, sicuramente con concezioni e approcci differenti rispetto agli studi effettuati negli ultimi anni, ma sicuramente volti ad identificare il problema e provare a fornire una, o più, possibili soluzioni. Risulta, tuttavia, difficile identificare quale sia l'origine del concetto di "stereotipo di genere" dato che le ricerche si sono fin da subito indirizzate agli stereotipi dati a donne e uomini, oltre che a gruppi di diverse etnie. I primi risultati su tali ricerche vengono dagli studi di Sheriffs e

Jarret nel 1953 e, qualche anno dopo, di Sheriffs e Mckee nel 1957, realizzati nel tentativo di individuare le credenze e le diverse caratteristiche attribuite a uomini e donne come stereotipi di genere.

Studi più recenti, invece, cercano di focalizzarsi su temi differenti: i contesti culturali in cui nascono gli stereotipi di genere, identificando somiglianze e differenze a livello di contenuto; le indicazioni fornite dall'Unione Europea su come trattare i *bias* di genere; gli stereotipi in età infantile, pre-adolescenziale, adolescenziale; i comportamenti causati dagli stereotipi di genere; l'evoluzione e il cambiamento del contenuto degli stereotipi di genere nel tempo.

Si è riscontrato come i media giochino un ruolo fondamentale nella trasmissione degli stereotipi, sia di genere che di altra natura. Immagini, articoli, post sui *social* e varie altre tecnologie utilizzate oggi sono strumenti che possono diffondere o consolidare idee, pensieri, opinioni stereotipate e al contempo modificare modelli che presentano questi *bias.* Ecco che l'impegno a ridurre i *bias* di genere è al centro del dibattito internazionale, tanto che la stessa comunità europea ha presentato *Una tabella di marcia per la parità tra uomini e donne*, tra i cui obiettivi principali spicca "eliminare gli stereotipi di genere nella società, nell'istruzione, nella formazione e nella cultura [...] nel mondo del lavoro [...] nei mezzi di comunicazione"[2].

1.1 Bias di genere nell'Intelligenza Artificiale

Smartphone, tablet, pc, assistenti vocali e IoT (Internet of Things) fanno sempre più uso dell'Intelligenza Artificale (AI): l'utilizzo di questa tecnologia rappresenta sicuramente una semplificazione di svariate operazioni quotidiane, tuttavia è sempre più evidente come l'AI porti con sé anche dei *bias* e tutte le conseguenze che ne derivano.

[2]https://www.europarl.europa.eu

L'AI viene definita da Marco Somalvico, ingegnere italiano specializzato nell'AI, come "una disciplina appartenente all'informatica che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana" (Somalvico, 1992)[3].

Per estrarre, tramite l'intelligenza artificiale, schemi e informazioni da dati grezzi o puliti e sistemati, ci si avvale di tecniche di apprendimento automatico (*machine learning*), invece per moderare o eliminare quasi completamente i comportamenti discriminatori all'interno degli algoritmi di *machine learning*, inclusi anche i dati utilizzati in questo elaborato e nei dati sui quali vengono allenati i modelli di intelligenza artificiale, si utilizzano tecniche di moderazione come il *debiasing*.

I *bias* nell'AI possono derivare da diverse fonti, come ad esempio preconcetti già presenti in coloro che la pensano e la progettano, possono essere dovuti a limiti tecnici derivati dalla progettazione del sistema stesso, oppure possono essere la rappresentazione del pensiero di chi produce e/o raccoglie i dati. Se i dati vengono raccolti da un sito nel quale tutti i post sono discriminatori verso le persone di colore e un modello viene allenato su questi dati, il modello stesso sarà affetto da un pregiudizio nei confronti delle persone di colore, semplicemente perché si è utilizzato un dato parziale per rappresentare un pensiero generale.

1.1.1 Bias nei dataset

La prima fonte di *bias* nel mondo dell'Intelligenza Artificiale è rappresentata dai dati utilizzati sia per allenare i modelli, sia per cercare di estrarre un risultato.

Ogni giorno noi stessi filtriamo grandi moli di dati, ma la quantità di dato a disposizione rispetto alla quantità di dato utilizzato è veramente irrisoria. Utilizzando questi pochi dati a disposizione si incorre nel rischio di avere *dataset* contenenti

[3]https://www.treccani.it/enciclopedia/intelligenza-artificiale_%28Enciclopedia-Italiana%29/

distorsioni della realtà, che portano i modelli allenati su questi *dataset* ad avere pregiudizi sistematici, ad esempio sull'età, sulla razza, sull'orientamento sessuale di un certo individuo o di un gruppo di individui.

Come è noto, gli strumenti di apprendimento automatico hanno la necessità di grandi quantità di dati di addestramento per produrre dei risultati validi. Avere, quindi, una numerosità ridotta di dati, dati distorti, o magari raccolti in maniera errata, comporta previsioni imprecise e risultati errati o qualitativamente bassi. Possiamo identificare diversi tipi di *bias* nei dati di addestramento:

- *Reporting bias*: quando è presente solo un numero ridotto di risultati all'interno del dataset che non ricopre l'intera popolazione. *Bias* che rispecchia la tendenza delle persone a non riportare tutte le informazioni disponibili. Tra i reporting bias si identificano:
  - *citation bias*: si verificano quando l'analisi si basa su studi che si trovano nelle citazioni di altri studi;
  - *language bias*: si verificano quando non si considerano i report non pubblicati nella propria lingua madre;
  - duplicate pubblication bias: si verificano quando alcuni studi vengono considerati maggiormente perché pubblicati in più luoghi;
  - *location bias*: si verificano quando alcuni report sono più difficili da trovare rispetto ad altri;
  - *publication bias*: si verificano quando studi con risultati positivi hanno maggiore probabilità di essere pubblicati rispetto ad altri con risultati negativi o risultati poco significativi;
  - *outcome reporting bias*: si verificano quando la segnalazione di alcuni esiti/attività è selettiva. Viene segnalato l'utile di un'azienda solo quando tale utile è positivo in un trimestre;
  - *time lag bias*: si verificano quando uno studio impiega anni per essere pubblicato.

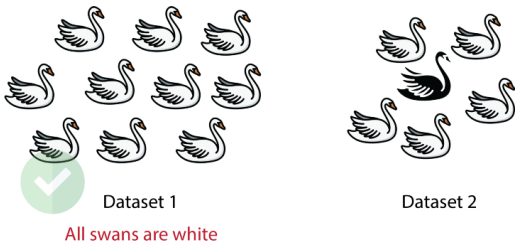

Figura 1.1: Overgeneralization Bias

- *Automation bias*: si verificano quando l'essere umano predilige i consigli e i risultati forniti da sistemi automatici, ignorando le informazioni contraddittorie date da sistemi non automatici nonostante la loro correttezza.
- *Selection bias*: si verificano quando i dati sono scelti in un modo che non rispecchia la loro distribuzione nel mondo reale. Spesso avviene perché non si effettua una randomizzazione consona nel processo di raccolta del dato. Alcune tipologie di questi *bias* sono:
  - *sampling bias*: si verificano quando il processo di randomizzazione del dato nel momento della sua raccolta non viene eseguito nel modo corretto;
  - *convergence bias*: si verifica quando i dati non vengono selezionati in modo rappresentativo. Per esempio, se in una raccolta dati da parte di un supermercato vengono intervistati solamente i clienti che hanno acquistato un certo tipo di prodotto e non quelli che ne hanno acquistato un altro, l'insieme di dati non rappresenterà il secondo gruppo di persone;
  - *participation bias*: si verificano quando i dati non sono rappresentativi poiché il numero di partecipanti al processo di raccolta dei dati è molto basso.
- *Overgeneralization bias*: si verifica quando si presume che i dati presenti nel proprio *dataset* siano simili a quelli presenti in un altro utilizzato per la stessa analisi, indipendentemente dalle loro dimensioni.
- *Group attribution bias*: si verifica quando si tende a stereotipare un gruppo basandosi solamente sulle azioni di alcuni individui all'interno dello stesso. Ci sono due tipologie di questo bias:

  - *in-group bias*: quando si fanno preferenze verso individui che appartengono al gruppo di cui si fa parte o con cui si hanno interessi comuni;
  - *out-group bias*: quando si creano degli stereotipi verso individui di gruppi ai quali non si appartiene.

- *Implicit bias*: si verificano quando si fanno delle ipotesi basandosi sulla propria esperienza personale, non applicabile in generale. Ad esempio, indicare il colore rosso come pericolo per una persona occidentale può risultare normale, mentre per una persona di origine cinese il colore rosso simboleggia fortuna, gioia e felicità. Una tipologia di *implicit bias* è il *confirmation bias* o *experimental bias* che si verifica quando si cercano informazioni che confermino o supportino le proprie convinzioni, idee o esperienze vissute.

1.1.2 Esempi di bias di genere nell'AI

OPTUM

OPTUM è un sistema, utilizzato negli ospedali statunitensi, che determina quali pazienti hanno necessità di ulteriori cure mediche. Nel 2019 è stato oggetto di diverse critiche poiché secondo alcuni studi l'algoritmo alla base discriminava i pazienti di colore rispetto ai pazienti caucasici.
L'algoritmo non era intenzionalmente razzista, ma escludeva specificamente la razza. Per identificare i pazienti che avrebbero beneficiato di ulteriori cure mediche l'algoritmo alla base utilizzava come metrica il costo sanitario futuro dei pazienti. Tale misura sicuramente non era una misura stereotipata, ma nemmeno neutra nell'identificare il bisogno di assistenza sanitaria. Analizzando i costi sostenuti dai pazienti di colore, è risultato che in media essi spendevano circa 1.800 dollari in meno rispetto a pazienti bianchi con lo stesso numero di patologie, di conseguenza l'algoritmo ha valutato pazienti bianchi come ugualmente a rischio di problemi di salute futuri, rispetto a pazienti di colore con molte più malattie.
Questo è un chiaro esempio di come dati che possono sembrare non affetti da *bias* di genere, in realtà siano affetti da pregiudizi sociali, culturali e istituzionali di lunga data, come ad esempio i costi dell'assistenza sanitaria.

COMPAS

Il *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) è l'algoritmo utilizzato in alcuni sistemi di giustizia statunitensi per prevedere la probabilità che un imputato commetta di nuovo un certo crimine. Non solo, l'algoritmo è studiato anche per identificare i bisogni dell'imputato in aree come l'occupazione, la disponibilità di alloggio ed eventuale abuso di sostanze stupefacenti. I dati forniti a COMPAS per determinare il grado di rischio e le eventuali esigenze, sono ottenuti elaborando i dati presenti nel fascicolo dell'imputato e le risposte fornite dallo stesso dopo un colloquio. Il risultato è un grafico a tre barre con punteggio da 1 a 10 rispettivamente per recidiva processuale, recidiva generale e recidiva violenta. Importante notare che COMPAS non fornisce una probabilità sulla recidiva individuale dell'imputato, bensì una previsione comparativa tra le informazioni ottenute dall'imputato e quelle di un gruppo di persone che hanno caratteristiche simili a quelle dell'imputato stesso.
Dopo diverse analisi è emerso che l'algoritmo è afflitto da un forte pregiudizio che porta ad avere il doppio di falsi positivi per recidiva se i trasgressori sono persone di colore (con una percentuale che si aggirava attorno al 45% circa) rispetto ai risultati ottenuti dai trasgressori caucasici (in questo caso la percentuale si attestava attorno al 23 %).
Anche in questo caso i bias erano dovuti a diversi elementi come i dati utilizzati, il modello scelto, la metodologia con la quale l'algoritmo era stato pensato e creato.
Gli studi su questo caso sono stati svariati, di particolare rilievo uno studio effettuato da Aliverti, Lum, Johndrow e Dunson dal titolo *Removing the influence of group variables in high-dimensional predictive modelling* in cui viene proposto, tra le altre cose, un modello per la rimozione delle informazioni sull'etnia da un set di dati.

CURRICULUM VITAE

Negli ultimi anni è sempre più utilizzata l'AI dalle aziende, soprattutto dalle multinazionali, per abbinare i CV ricevuti alle "posizioni aperte". Molto spesso, però, gli algoritmi di *job recruiting* utilizzati presentano *bias* di genere e, in particolare, sfavoriscono l'occupazione femminile.

Un esempio di ciò è quanto successo a Linkedin, che ha trovato i propri algoritmi di raccomandazione fortemente distorti, tanto che nelle posizioni disponibili, i candidati di sesso maschile erano maggiormente presenti e significativamente più propensi alla ricerca di nuove opportunità lavorative rispetto alle candidate di sesso femminile. Appena trovata questa falla nel proprio sistema, Linkedin si è subito attivato per realizzare un nuovo algoritmo di intelligenza artificiale che contrastasse tale *bias*. Questo ha fatto sì che molte piattaforme di ricerca del lavoro come CareerBuilder, ZipRecruiter e Monster, si attivassero per utilizzare approcci differenti al fine di contrastare l'insorgenza di gender *bias*. Linkedin, dopo quanto successo, ha escluso nome, età, sesso e razza di un candidato dalle caratteristiche selezionate per la valutazione di un CV, in quanto possono aumentare distorsioni nel processo di selezione.
Altro caso di studio è stato quello di Amazon che tra il 2014 e il 2017, nella selezione dei candidati, avrebbe penalizzato i CV composti da parole come donna o da nomi di college di genere femminile.

Interessante risulta essere la strada intrapresa da CareerBuilder, che focalizza la propria politica su come utilizzare i dati raccolti dal servizio proponendo alcune modalità per eliminare eventuali pregiudizi di genere dalle offerte di lavoro, consigliando, ad esempio, l'utilizzo di un determinato linguaggio all'interno delle proposte di lavoro e garantendo così un *recruiting* non penalizzante verso uno dei due sessi.

1.2 Bias nei documenti testuali

Come riportano molti studi il linguaggio ha un forte impatto su quello che si percepisce, su quello che si pensa e sulla definizione dei ruoli di genere (Doughman, Khreich, 2022; Doughman, Khreich, El Gharib, Wiss, Berjawi, 2021, p.34-44; Sun, Gaut, Tang, Huang, ElSherief, Zhao, Mirza, Belding, Chang, Wang, 2019). Di conseguenza, cercare di avere un linguaggio quanto più inclusivo possibile è diventata una sfida che negli ultimi anni è stata portata avanti da diverse realtà e movimenti in tutto il mondo. Il linguaggio non va inteso solamente come quello parlato, anche

il linguaggio scritto, infatti, ha una notevole importanza nel cercare di eliminare i *bias* di genere diffusi nella società moderna.

Proprio per questo, diverse tecniche, come l'utilizzo dei WEs e successivamente dei Sentence Embeddings, affiancate dalla Text Analysis e dalla Sentimenti Analysis, sono state sviluppate, studiate e migliorate per cercare di ridurre al minimo i *gender bias* all'interno dei documenti testuali.

Già nel 1975 Bodine scoprì che l'uso generico del pronome *he*, nella lingua inglese, proviene da una visione del mondo androcentrica diffusa tra i grammatici del XVIII secolo. Questi pregiudizi nel linguaggio comportano un continuo feedback negativo causato dall'impatto diretto del linguaggio su quello che una persona percepisce (Boroditsky, 2011, p.62-65). L'uso ricorrente di pregiudizi nel linguaggio porta ad una percezione del mondo distorta che si ripercuote poi nella scelta lessicale di ogni persona. Questo viene notevolmente amplificato, come accennato in precedenza, dai sistemi di AI.

Potremmo definire il *bias* nei documenti testuali come una rappresentazione che esclude, generalizza o crea un pregiudizio su un unico genere basandosi su vari stereotipi della società.

La ricerca, in questo ambito, si è indirizzata in particolare nel cercare di eliminare i *bias* di genere nei CV e nei libri di testo utilizzati nelle scuole. Questo secondo punto è molto sentito dalla comunità scientifica poiché crescere generazioni libere dai *gender bias* potrebbe assicurare un futuro meno legato a stereotipi e pregiudizi di genere. L'analisi del contenuto dei libri di testo ha mostrato come le donne occupino uno stato di subordinazione all'interno della società, in quanto il ruolo che ricoprono nella maggior parte dei casi è poco significativo e legato al lavoro di casalinga. Gli uomini invece svolgono ruoli principali nel 75% dei casi e sono indicati come medici, scienziati, presidenti, venditori, ma anche agricoltori e professionisti.

Entrando più nel dettagli di alcuni casi riportati in letteratura, studi sui libri di testo cinesi per l'infanzia e la primaria citati nella GMR 2008 dell'EFA (European Financial Advisor), hanno mostrato come i maschi fossero rappresentati in modo sproporzionato e le femmine apparissero frequentemente solo nei materiali di lettura per i bambini molto piccoli. La percentuale di personaggi maschili è passata dal 48% nei libri per bambini di 4 anni al 61% in quelli per bambini di 6 anni. Nei testi di studi sociali scienziati e soldati erano rappresentati come maschi, al contrario gli insegnanti e tre quarti del personale di servizio erano donne. Le donne rappresentavano solo un quinto dei personaggi storici nei dodici volumi dei libri di testo primari e apparivano spente e prive di vita rispetto ai maschi più vivaci.

Mediamente in India, oltre la metà delle illustrazioni dei libri di testo di scuola primaria di inglese, hindi, scienze, matematica e studi sociali raffigurava solo maschi e appena il 6% raffigurava esclusivamente femmine. All'interno dei libri di matematica utilizzati nelle scuole primarie, le attività riguardanti situazioni commerciali, occupazionali e di marketing erano svolte da uomini, e nessuna donna è stata raffigurata come dirigente, negoziante, commerciante o ingegnere. Un lavoro interessante riguardante questa tematica è *Gender stereotypes in education: Development, consequences, and interventions* (Kollmayer, Kollmayer, Spiel, 2018, p.361-377)

La ricerca ha mostrato come alla fine degli anni 2000, in paesi come il Camerun, la Costa d'Avorio, il Togo e la Tunisia, la proporzione di personaggi femminili all'interno dei libri di testo di matematica rispetto a quelli maschili fosse di appena il 30% . Entrambi i generi venivano generalmente rappresentati con ruoli professionali o domestici fortemente stereotipati. Mentre i ragazzi e gli uomini avevano facevano cose impressionanti, nobili, eccitanti e divertenti, le donne venivano ritratte come lavoratrici domestiche, accomodanti, accudenti e le ragazze come conformiste passive.

Capitolo 2

Natural Language Processing

Il Natural Language Processing (NLP) è quell'insieme di tecniche che utilizzano algoritmi di Intelligenza Artificiale in grado di rappresentare, analizzare e comprendere il linguaggio naturale. Gli ambiti di utilizzo sono i più svariati, dalla comprensione del contenuto, alla traduzione, nonché la produzione di un testo in maniera automatizzata, basandosi su dati o documenti forniti in input. Nonostante le tecniche di ML e Deep Learning abbiano portato notevoli innovazioni e miglioramenti degli algoritmi di NLP, per quanto riguarda la lingua italiana le tecniche ad oggi presenti non risultano essere così performanti come in altre lingue, tipo l'inglese. Questo perché la lingua italiana è caratterizzata da molti modi di dire, espressioni gergali ed è fortemente influenzata da numerosi dialetti.

Un aspetto importante dell'NLP è sicuramente la linguistica computazionale, ovvero lo studio e la costruzione di sistemi informatici che eseguono un'analisi e conseguentemente un'elaborazione del linguaggio naturale, che si focalizza sullo studio del funzionamento del linguaggio in modo da generare programmi eseguibili dalle macchine.
Principalmente l'NLP si occupa di testi, visti come sequenze di parole, che in una certa lingua esprimono un certo concetto o messaggio, come ad esempio documenti aziendali, file di log, tweet, post nei social, pagine web, CV e molti altri. L'NLP non si occupa solamente di linguaggio scritto, infatti sempre di più la ricerca si sta spostando verso l'elaborazione del linguaggio parlato. Il dialogo uomo-macchina

racchiude diverse componenti come la fonetica, la fonologia, la morfologia, la sintassi, la semantica, la pragmatica e il senso del discorso nel suo complesso.

Sono numerosi i task di NLP che cercano di automatizzare compiti semplici in queste aree:

- riconoscimento della lingua
- analisi del sentiment
- analisi semantica
- scomposizione della frase in unità elementari

In questo modo si possono identificare delle macro aree di task di NLP:

- *Language Translation*: traduzione del linguaggio di un documento, scegliendo per ogni parola il significato migliore in base al contesto;
- *Text Generation*: generazione automatica di un testo;
- *Text Analysis*: analisi di un testo ed individuazione di elementi principali;
- *Smart Search*: recupero di documenti compatibili con una ricerca posta in linguaggio naturale;
- *Sentiment Analysis*: determinazione del livello di sentiment di un documento (ad esempio una recensione negativa o una positiva). In questo ambito è interessante lo studio *More than Bags of Words: Sentiment Analysis with Word Embeddings* (Rudkowsky, Haselmayer, Wastian, Jenny, Emrich,Sedlmair, 2018, p.140-157);
- *Text Classification*: interpretazione di un testo per determinarne la categoria (nel caso delle mail ad esempio, per trovare mail di spam);
- *Automatic Summarization*: produzione di un riassunto di uno o più documenti testuali;

- *Intent Monitoring*: comprensione di un testo per prevedere comportamenti futuri (ad esempio la volontà di acquisto di un cliente).

Sempre più spesso le aziende sono interessate a soluzioni di NLP per risolvere diversi problemi, per generare e cogliere diverse opportunità di business o per automatizzare processi o attività routinarie:

- analisi delle email aziendali, così da classificare i messaggi in entrata ed eliminare eventuali messaggi indesiderati o suddividerli per categoria, come ad esempio permette di fare Google classificando le mail come importanti, promozioni, ecc.;

- estrazione di informazioni da documenti di *governance*, come report o procedure garantendo così una consultazione più rapida ed efficace;

- progetti per l'analisi di documenti amministrativi, come fatture o contratti, e soluzioni per la gestione delle comunicazioni interne all'azione come la gestione delle mail ricevute dal proprio portale di help-desk;

- analisi di post sui Social, ad esempio per comprendere il *sentiment* degli utenti su un determinato argomento e modificare la strategia di social marketing dell'azienda;

- algoritmi per la comprensione di *query* di ricerca nel web con conseguente re indirizzamento alle pagine più adatte all'argomento cercato;

- algoritmi per la *detection* di fake news su portali giornalistici, social o pagine web;

- classificazione del livello di privacy dei documenti;

- *data entry* automatico con estrazione delle principali entità dal testo.

2.1 Word Embeddings

I *Word Embeddings*, (WEs), sono modelli che fanno parte di quella branchia dell'informatica che si occupa di Natural Language Processing (NLP) e che mappano parole in uno spazio vettoriale multi-dimensionale:

$$V \to \Re^D : w \mapsto \vec{w} \tag{2.1}$$

Questi vettori permettono di memorizzare informazioni semantiche e sintattiche delle parole mappate, in maniera tale che vettori simili, ovvero riconosciuti semanticamente simili, siano più vicini tra loro rispetto a vettori rappresentanti parole che non rientrano negli stessi concetti linguistici. Un esempio di questa correlazione fra vettori (Figura 2.1) è data dall'analogia:

$$man : king = woman : queen \tag{2.2}$$

Questo si basa sul fatto che il vettore differenza *man* - *woman* ha la stessa direzione di quello *king* - *queen*.

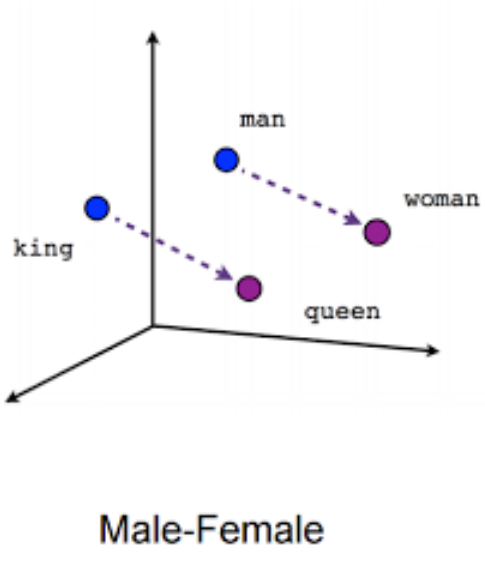

Figura 2.1: Esempio correlazione tra vettori

Tra i vari modi utilizzati per determinare questa correlazione semantica tra vettori, il più utilizzato è sicuramente la metrica della *cosine similarity*. Tale metrica

si basa sul calcolo del coseno dell'angolo tra due vettori: più il valore è vicino ad 1 più i due vettori saranno simili.

$$cosineSimilarity(\vec{u}, \vec{v}) = cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{||u|| \cdot ||v||} \tag{2.3}$$

Di seguito alcuni esempi di quanto appena descritto. I vettori sono recuperati dalla libreria FastText.

Coppia	**Score**
Uomo - Donna	0.66293514
Re - Regina	0.4141101
Uomo - Re	0.083193354
Donna - Regina	0.46889493

Ci sono diversi metodi utilizzati per generare i WEs, in particolare vengono utilizzate le reti neurali, i modelli probabilistici e altre tecniche basate sulla matrice di co-occorrenza delle parole. Generalmente le reti neurali sono lo strumento più utilizzato per il calcolo dei WE e vengono allenate su grandi corpus testuali come Common Crawl[1], Wikipedia dumps[2] e molti altri. A questi metodi si affiancano vari algoritmi per l'apprendimento, come Word2Vec di Tomas Mikolov o GloVe della Standford University. Librerie come Gensim o Deeplearning4j consentono di implementare questi algoritmi e sfruttarli in altre applicazioni.
Di seguito una breve spiegazione di come W2V e GloVe sono implementati.

2.2 Word2Vec

Word2Vec è un modello, sviluppato da Tomas Mikolov nel 2013 in Google che permette di costruire il vettore relativo di una parola (Mikolov, Chen, Corrado, Dean, 2013). W2V utilizza le *neural networks* per generare tali vettori ed in particolare può utilizzare uno dei seguenti metodi:

- CBOW (Common Bag Of Words)

[1] https://commoncrawl.org/
[2] https://dumps.wikimedia.org/enwiki/

- Skip Gram

Prima di procedere con una breve spiegazione dei due metodi è utile notare che W2V sfrutta una tecnica spesso utilizzata nel ML: fare il training di una semplice *neural network*, con un singolo *hidden layer*, per completare un certo task che poi non sarà quello effettivamente utilizzato dalla *neural network*. Pensiamo al caso della compressione di un vettore: la rete viene addestrata per comprimere un certo vettore in input e poi decomprimere lo stesso per ottenere il vettore di partenza. Una volta che la rete è allenata, vengono mantenuti solamente l'*input layer* e l'*hidden layer*, scartando il *layer* di output che non risulta necessario al *task* di compressione del vettore. Allo stesso modo quello di cui necessita W2V è imparare a pesare nella maniera corretta il vettore nell'*hidden layer*, utilizzandoli come rappresentazione vettoriale (WEs) della parola.

2.2.1 Skip Gram

Nel modello Skip Gram, per addestrare la rete neurale, vengono passate coppie di parole per svolgere il "finto" *task*, ovvero data una parola in input viene restituita la distribuzione di probabilità delle parole ad essa vicine, ovviamente in riferimento alla sua posizione nel testo. Ad esempio, presa la frase: "vado al mare a nuotare", e data in input la parola "mare", le parole ad essa vicine saranno [vado, al, a, nuotare]. Questo come detto è il "finto" *task*, e quello che il modello restituirà è solamente il vettore di pesi calcolati nell'*hidden layer*, ovvero la rappresentazione vettoriale della parola.

2.2.2 CBOW

Il modello CBOW fa esattamente l'opposto, ovvero, partendo dal contesto, cerca di ricavare la parola principale. Come nel caso precedente, quello che poi verrà restituito sarà il vettore dei pesi associato all'*hidden layer*, che rappresenta così la parola cercata.

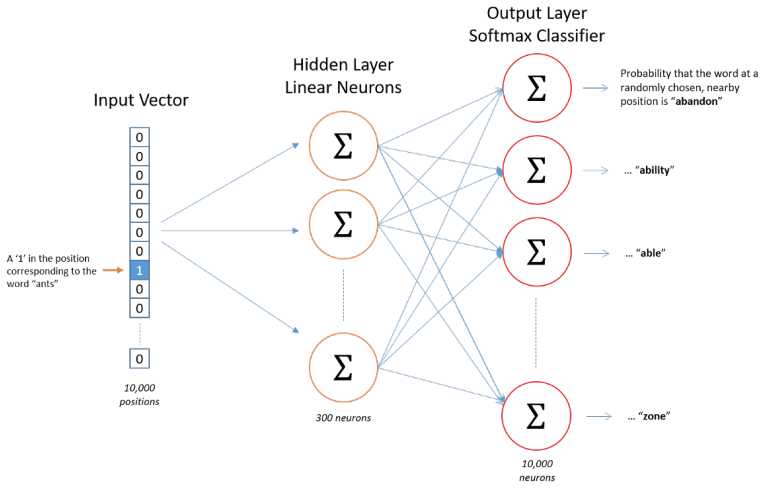

Figura 2.2: Skip-Gram model

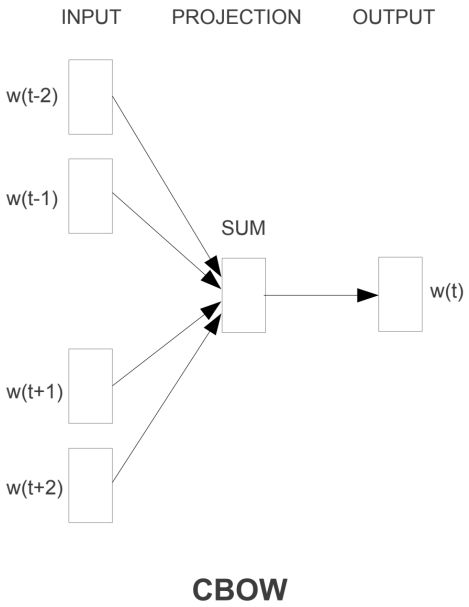

Figura 2.3: CBOW model

2.3 Glove

La prima versione di GloVe è stata rilasciata da Jeffrey Pennington nell'Agosto del 2014. Si tratta di un algoritmo di apprendimento automatico, non supervisionato, usato per rappresentare le parole come vettori. L'addestramento viene effettuato usando statistiche aggregate di co-occorrenza parola-parola all'interno di un corpus e tali rappresentazioni forniscono interessanti sotto-strutture lineari dello spazio vettoriale delle parole.

Un singolo valore scalare viene prodotto dalle metriche di somiglianza utilizzate per trovare il *nearest neighbor*, quantificando quanto le due parole in esame siano collegate. Talvolta questo semplice risultato può essere un problema dato che spesso due parole presentano relazioni più complesse di quelle che possono essere rappresentate con un singolo numero. Prendiamo l'esempio delle parole uomo/donna: tali parole possono essere considerate simili in relazione alla descrizione del genere umano, mentre se consideriamo la direzione di genere queste due parole indicano versi opposti lungo tale direzione e sono, di conseguenza, considerate due parole opposte. Proprio per risolvere questo problema GloVe è stato progettato affinché le differenze vettoriali possano catturare il significato, specificato dalla giustapposizione di due parole.

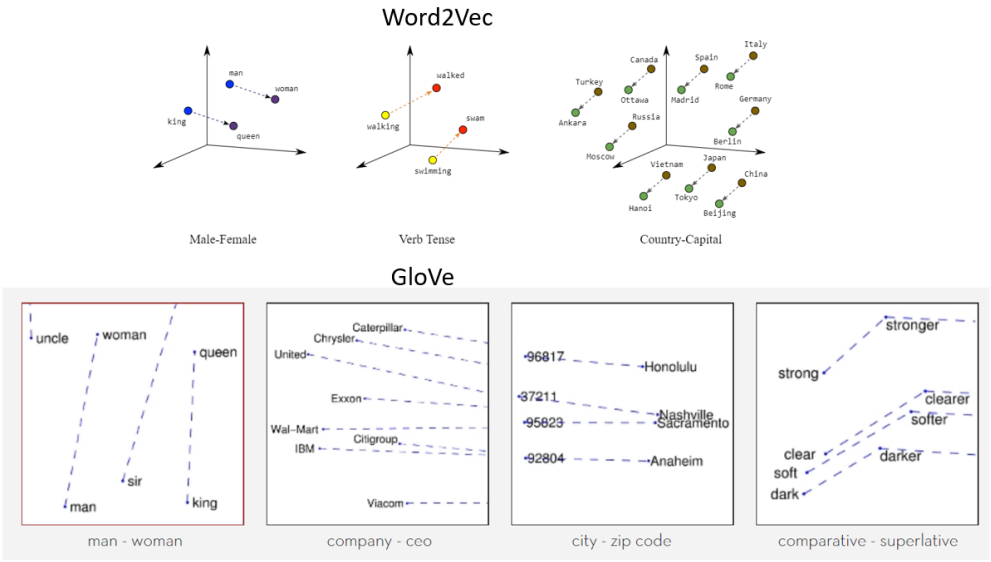

Figura 2.4: GloVe vs W2V

Come accennato in precedenza, i WEs utilizzano spazi vettoriali multi-dimensionali, dove l'alto numero di dimensioni aiuta a catturare più informazioni possibili. Il modello utilizzato in questo studio utilizza vettori 300-dim presi dalla libreria FastText. In generale, un concetto importante da tenere presente è che le reti neurali sono degli approssimatori universali, di conseguenza più si aumenta il numero di dimensioni, più ci si avvicina ad un risultato simile alla realtà. Va considerato, però, che ad un certo punto il costo di una nuova iterazione nel calcolo dimensionale risulta essere maggiore rispetto al beneficio che porta in termini di ottimizzazione del risultato. Di conseguenza trovare la giusta dimensione dei vettori è uno dei punti chiave nella definizione della rete neurale stessa.

Strumenti come l'*analisi PCA* o il *t-distributed stochastic neighbor embedding* sono usati per ridurre lo spazio vettoriale e consentire una più semplice rappresentazione dei vettori in spazi 2D o 3D.

Capitolo 3

Word Embeddings: alcuni casi di studio

Nel seguente capito verranno analizzati i due paper principali, utilizzati come guida sia per la parte di studio, sia per la parte di analisi dei dati. Tali paper sono: *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (Bolukbasi, Chang, Zou, Saligrama, Kalai, 2016) e *Gender Bias in Italian Word Embeddings* (Biasion, Fabris, Silvello, Susto, 2020).

3.1 Uomo sta a programmatore come donna sta a y

Bolukbasi (2016) analizza i pregiudizi di genere nell'apprendimento automatico come risultato dell'utilizzo di dati di addestramento distorti e propone una soluzione per eliminare il *bias* dal modello.

Gli autori utilizzano un *word embedding model* per dimostrare i *gender bias* all'interno dei dati di stima. Il modello è allenato su articoli di Google News e la rappresentazione vettoriale di tutte le parole è fatta tramite vettori in 300 dimensioni.
Attraverso la *cosine similarity* è possibile trovare quanto due vettori siano vicini tra loro e, di conseguenza, capire se due parole sono semanticamente simili.

La discussione trattata nel paper si basa proprio sulla necessità di individuare le analogie tra le parole utilizzando la loro rappresentazione vettoriale, ad esempio trovare l'analogia "Man is to x as Woman is to y". Usando l'algebra lineare è possibile calcolare la differenza tra il vettore Man e il vettore Woman e trovare la coppia di vettori la cui differenza è più vicina a quella trovata. La coppia di vettori più vicina alla differenza vettoriale Man-Woman sarà la migliore analogia per completare la frase "Man is to x as Woman is to y".

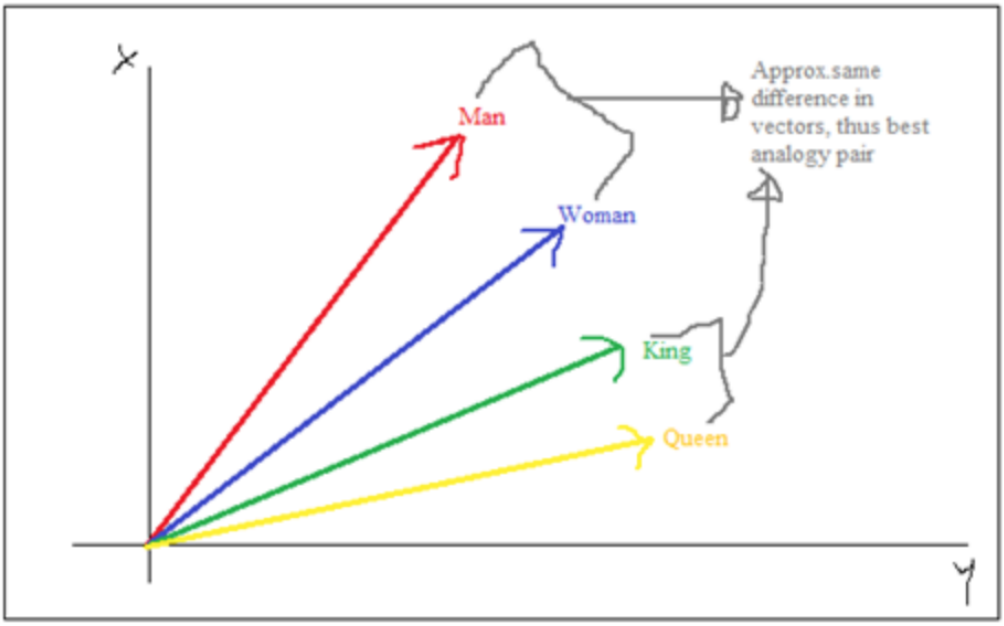

Figura 3.1: Vector analogies

L'immagine riportata sopra mostra come la miglior analogia sia "Man is to King as Woman is to Queen". Questo dimostra come i WEs siano uno strumento molto potente per catturare una varietà di relazioni all'interno di un corpus utilizzando della semplice aritmetica sui vettori.

Gli autori mostrano come il sessismo implicito contenuto nei dati di allenamento venga amplificato nei modelli WEs. Ad esempio, analogie sessiste ritornate dal modello sono:

- Man is to computer programmer as woman is to homemaker.
- Father is to doctor as mother is to nurse.

Questo mostra come i dati negli articoli di Google News contengano gender *bias* e di come i WEs li amplifichino.

3.1.1 Valutazione dei bias

Una parola "*gender-specific*" è una parola che è appropriata per il genere e non mostra *bias*. Ad esempio, la parola "Father" è *gender male* e la parola "Mother" è *gender female*, di conseguenza Father-Mother è una coppia di parole *gender-specific*.

Per quantificare i *bias*, gli autori comparano una *word vector* ad una coppia di gender-specific words. Consideriamo, ad esempio, il vettore della parola "nurse" e la coppia gender-specific "Father-Mother". La parola "nurse" dovrebbe essere gender-neutral.

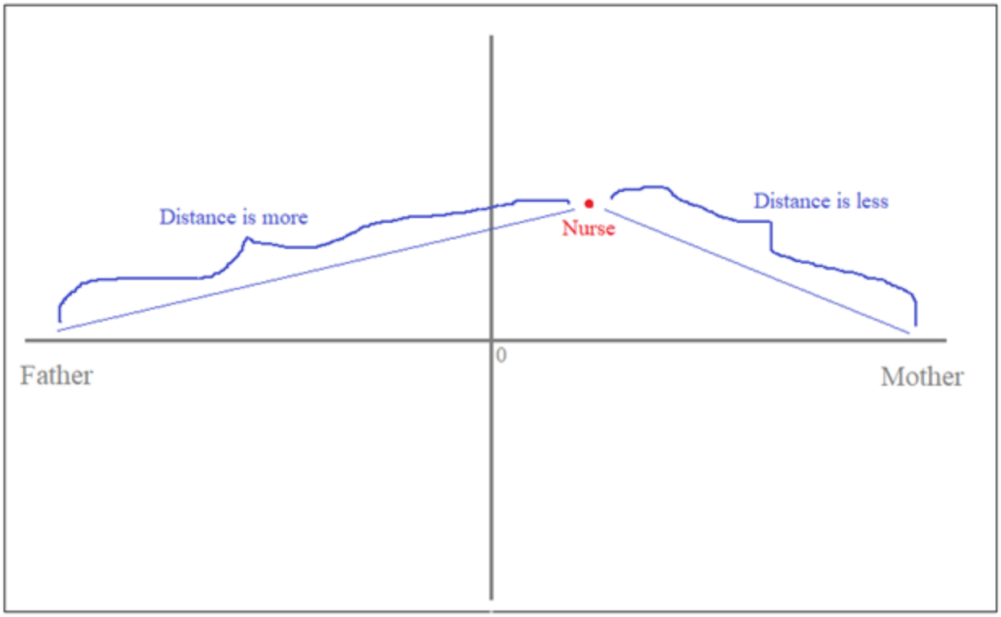

Figura 3.2: Nurse evaluation

Come si può vedere dal grafico riportato, la distanza Nurse-Mother è più piccola rispetto alla distanza Nurse-Father, questo implica che la parola Nurse sia più femminile, ma ciò non è corretto poiché un "nurse" (infermiere/a) può essere un maschio o una femmina.

Nel paper vengono identificati due tipi di gender *bias*:

- *Direct bias*: in questo tipo di *bias*, le parole *gender-neutral* sono proiettate sui *gender axis* e dovrebbero essere idealmente neutre. Se, invece, c'è una tendenza verso uno dei due assi (maschile o femminile) allora c'è un *bias*.

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Figura 3.3: Dircet bias example

- *Indirect bias*: In questo tipo di *gender bias*, le parole *gender neutral* sono proiettate lungo assi "estremi" delle professioni. Ad esempio, *softball* è considerato una professione estremamente femminile, mentre *football* una professione estremamente maschile. Idealmente, parole *gender neutral* dovrebbero essere nel mezzo degli assi *softball-football*. Se, invece, tende verso uno dei due assi allora c'è un *indirect bias*.

***softball* extreme**
1. pitcher
2. bookkeeper
3. receptionist
4. registered nurse
5. waitress
***football* extreme**
1. footballer
2. businessman
3. pundit
4. maestro
5. cleric

Figura 3.4: Indirect bias example

Per verificare i *bias* sono stati impiegati gli *US-based Amazon mechanical turks* per definire un *bias rating* sulle parole o sulle analogie. *Amazon Mechanical Turk* (MTurk) è un servizio internet di *crowdsourcing* che permette a persone come programmatori, ricercatori, o chiunque ne abbia necessità, chiamati *requester*, di coordinare l'uso di intelligenze umane per eseguire compiti che i computer, ad oggi, non sono in grado di eseguire.

3.1.2 Procedura di debiasing

Per effettuare il *debiasing* gli autori hanno seguito il seguente modello: per prima cosa è stato individuato un piccolo insieme di *gender-specific words* come "he, she, man, woman, ..." da utilizzare come *seed words*. Con queste parole è stato creato un *Support-Vector Machines* (SVM) *classifier* per ottenere tutte le *gender specific words*. Tramite le parole *gender specific* è stato effettuato un complemento con il corpus per trovare le parole *gender neutral* come "programmer, homemaker, gardner, ...". In conclusione, si può trovare una separazione tra le parole *gender specific* e quelle *gender neutral* come si può vedere in Figura 3.5.

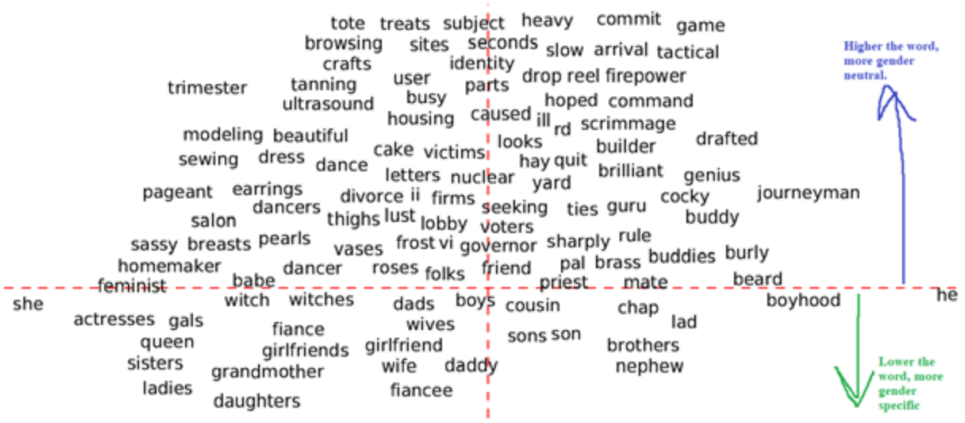

Figura 3.5: Words separation

Viene, quindi, scelta una coppia di parole per essere l'asse x. Come asse è stata scelta la coppia "she-he" e le altre parole proiettate su questi *gender axis*. Tale step viene identificato nel paper come *"identify gender subspace"*, ovvero identificare la

direzione degli *embeddings* che cattura i bias. Si può effettuare il *debiasing* di parole *gender neutral* in due modi: *hard debias* (neutralize e equalize) o *soft debias* (soften). *Hard debias* è una misura estrema per la quale nessuna traccia di genere nelle parole *gender neutral* rimane dopo l'applicazione. Il *soft debias* mantiene il genere nelle parole *gender neutral* in base ad un iperparametro specificato.

Nella parte di *neutralize* vengono proiettati tutti i termini *gender neutral* lungo l'asse y, così da eliminare il *gender bias*.

Figura 3.6: Neutralize step

Nella parte di *equalize* si rendono le parole *gender neutral* equidistanti da tutte le parole in ogni *equality set*, come grandmother, grandfather, guy, gal. Questo significa che la parola *babysitter* sarà equidistante dagli assi *grandmother-grandfather* e equidistante dagli assi *guy-gal*.

Il paper, infine, specifica che il modello di *debiasing* riduce le analogie stereotipate e preserva l'usabilità degli *embeddings*.

3.2 Italian Word Embeddings

Quanto fatto da Biasion, Fabris, Silvello, Susto (2020) cerca di valutare se i *gender bias* presenti negli *italian* WEs codifichino stereotipi di genere studiati dalla psicologia sociale o che spesso si ritrovano nel mondo del lavoro.

I WEs relativi alla lingua italiana sono stati sviluppati e analizzati basandosi principalmente su quanto fatto dagli studi per i WEs inglesi (Berardi, Esuli, Marcheggiani, 2015; Bojanowski, Grave, Joulin, Mikolov, 2016; Tripodi, Pira, 2017), anche se sono rimasti fondamentalmente più indietro rispetto ai corrispondenti in-

glesi.

La prima parte del lavoro svolto dagli autori si basa sulla determinazione di sottospazi vettoriali che possano identificare direzioni di genere sulle quali proiettare i vari WEs per calcolare il corrispettivo *bias* di genere. In particolare, sono state definite sei coppie di parole di genere opposto:

- lui, lei;
- uomo, donna;
- padre, madre;
- marito, moglie;
- fratello, sorella;
- maschio, femmina.

Sono state evitate coppie di parole con la stessa radice, come figlio-figlia, per evitare l'intreccio con il genere grammaticale. Successivamente è stata applicata un'analisi *PCA* sui sei vettori differenza ottenuti sottraendo i corrispondenti vettori di ogni coppia l'uno dall'altro. La componente principale è risultata essere la prima con il 57% di varianza ed è, quindi, stata presa come *main gender direction*. Sostanzialmente questa procedura ha permesso agli autori di determinare una direzione che identificasse tutte le altre.
Il calcolo del *gender score* è stato effettuato utilizzando la seguente formula:

$$s_g(w) = \mathbf{w} \cdot \mathbf{g}/(|\mathbf{w}||\mathbf{g}|) \tag{3.1}$$

In particolare:

- $\mathbf{w}$ = vettore corrispondente alla parole w
- $\mathbf{g}$ = vettore direzione considerata

Grazie a questa formula gli autori hanno identificato l'associazione di w lungo la direzione scelta. Se il punteggio ottenuto era altamente positivo, allora la parola

scelta era più vicina alla componente maschile della coppia che identifica la direzione, se altamente negativa, invece, era più vicina alla componente femminile.
Tramite un'analisi *WEAT* (*Word Embedding Association Test*) è stata determinata la correlazione implicita tra categorie e concetti. Per un approfondimento su questo è possibile consultare il paragrafo 3.1.2 dell'articolo *Gender Bias in Italian Word Embeddings* (Biasion, Fabris, Silvello, Susto, 2020).
Infine, gli autori hanno cercato di gestire quello che viene definito *grammatical gender*, ovvero il fatto che la lingua italiana è una lingua genderizzata nella quale ad ogni sostantivo viene assegnato un genere grammaticale. Proprio per questo motivo, gli autori cercano di mitigarne gli effetti con la seguente formula:

$$s_{meang}(w) = (s_g(w_f) + s_g(w_m)) \tag{3.2}$$

Per le parole che ammettono sia la versione maschile che femminile (es: dottore, dottoressa) viene calcolata la media degli *score* di entrambe le parole rispetto alla direzione considerata, dando così ad entrambe le versioni della parola lo stesso peso.

Gli autori hanno effettuato, poi, vari esperimenti utilizzando parole che facevano riferimento ad occupazioni lavorative, scienza ed arte, carriera e famiglia.
Nei riferimenti alle categorie "scienza ed arte" e "carriera e famiglia", gli autori hanno effettuato un'analisi WEAT: nel primo caso hanno trovato che non ci fossero particolari associazioni stereotipate e hanno ipotizzato che tale evento fosse legato al genere grammaticale femminile delle parole target collegate all'ambiente scientifico; nel secondo caso, invece, è risultato che associazioni stereotipate fossero più marcate nel *dataset wiki*.

Gli autori hanno, infine, proposto un metodo per mitigare l'effetto del *grammatical gender* che si può approfondire nel paragrafo 5.4 di *Gender Bias in Italian Word Embeddings* (Biasion, Fabris, Silvello, Susto, 2020).

Quanto evidenziato dall'articolo è che i WEs italiani sembrano avere meno "potenziale" rispetto ai loro corrispettivi inglesi, soprattutto a causa della forte presenza di *grammatical gender* che porta con sé differenti problemi. Come prova di ciò, gli

autori riportano l'esempio della ricerca del lavoro, nella quale si potrebbe ritrovare che il genere maschile sia quello predefinito nelle *query* di analisi, ad esempio si ricerca la parola "psicologo", favorendo quindi i CV maschili rispetto a quelli femminili e mettendo, di conseguenza, le donne in una posizione di svantaggio.

Capitolo 4

Determinazione del gender level di un documento

Come accennato nel capitolo 3, i paper che hanno guidato questo lavoro sono stati *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (Bolukbasi, Chang, Zou, Saligrama, Kalai, 2016) e *Gender Bias in Italian Word Embeddings* (Biasion, Fabris, Silvello, Susto, 2020). In particolare, *Gender Bias in Italian Word Embeddings* (Biasion, Fabris, Silvello, Susto, 2020) è stato preso come modello di riferimento per la parte iniziale di studio e analisi dei WEs.

Si possono individuare tre stadi nel lavoro svolto:

- analisi dei word embeddings
- analisi degli n-grams
- analisi dei sentence embeddings

Questi tre livelli possono essere visti come una naturale successione l'uno dell'altro: si è partito con l'analisi della singola parola, non tenendo conto del contesto della stessa, semplicemente proiettandola rispetto ad una direzione prestabilita e calcolando lo *score* relativo. Si è passato, in seguito, all'analisi di gruppi di due, tre, quattro, cinque, sei parole in sequenza tra loro, andando a mediare lo *score* relativo di ogni singola parola nella coppia/trio ecc. Infine, si è provata a determinare una

percentuale di importanza di ogni singola parola all'interno di una frase. Lo *score* finale si è ottenuto moltiplicando lo *score* di ogni singola parola per la relativa percentuale di importanza all'interno frase.

4.1 Word Embeddings analisys

Per poter analizzare i WEs italiani è stato necessario scegliere un modello che restituisse, data una parola, il relativo vettore. La scelta è ricaduta sul modello *cc.it.300* della libreria FastText [1] . FastText è una libreria *open source* che consente a chi la utilizza di svolgere *text representation* e *text classification*. In particolare, il modello utilizzato è stato addestrato utilizzando CBOW con pesi di posizione, in dimensione 300, con n-grammi di caratteri di lunghezza 5, una finestra di dimensione 5 e 10 negativi.

Una volta scelto il modello, sono state definite le coppie di parole *gender neutral* per la determinazione delle direzioni gender lungo le quali proiettare i WEs. Come per Biasion, Fabris, Silvello e Susto (2020), le coppie risultano essere: [('lui', 'lei'), ('uomo', 'donna'), ('padre', 'madre'), ('marito', 'moglie'), ('fratello', 'sorella'), ('maschio', 'femmina')]. Per ogni parola nella coppia è stato determinato il corrispondente WEs e definito il vettore differenza risultante, ottenuto come differenza vettoriale tra la componente maschile e quella femminile.

Dopo aver effettuato un'analisi PCA sulle sei direzioni principali, al fine di determinare una direzione che le identificasse tutte, è risultato che la prima componente delle sei avesse una varianza del 39% e fosse dominante sulle altre. Confrontando, però, quanto ottenuto con i risultati forniti da Biasion, Fabris, Silvello e Susto (2020), che riscontravano una varianza del 57% sulla prima componente, si è deciso di sommare alla prima componente anche la seconda (varianza del 20%), ottenendo così una varianza complessiva del 58%, valore ritenuto più adeguato al caso di studio. Per avere un vettore risultante si è calcolata la somma vettoriale fra il vettore

[1] https://fasttext.cc

relativo alla prima componente trovata e quello relativo alla seconda, ottenendo così un unico vettore.

Nella continuazione dello studio si è scelto di confrontare la direzione lui-lei con la direzione ottenuta dalla somma delle due componenti principali ma di mantenere come direzione unica quella rappresentata dal vettore differenza lui-lei, che risultava essere un buon metro di paragone con quanto ottenuto da Biasion, Fabris, Silvello e Susto (2020).

Lo studio dei WEs italiani si è concentrato sull'analisi della proiezioni di parole inerenti ad occupazioni lavorative suddivise in due categorie:

- lavori ungendered
- lavori gendered

Nella prima categoria rientravano le seguenti parole, che risultavano essere utilizzate sia per identificare un lavoratore uomo, che una lavoratrice donna, come: elettricista, camionista, ingegnere, commercialista, notaio, architetto, dentista, medico, giornalista , barista , igienista, farmacista, preside, dietista, badante, insegnante. Nella seconda categoria, invece, vi erano parole che ammettono sia la forma maschile, sia la forma femminile: calzolaio, agrotecnico, geologo, avvocato, albergatore, veterinario, filosofo, zoologo, biologo, professore, psicologo, ostetrico, maestro, le cui controparti femminili sono: calzolaia, agrotecnica, geologa, avvocatessa, albergatrice, veterinaria, filosofa, zoologa, biologa, professoressa, psicologa, ostetrica, maestra.
Per poter correlare il *gender score* del WEs relativi alle professioni è stato necessario recuperare le percentuali di occupazione di uomini e donne nei lavori sopra elencati. Questo è stato fatto cercando dati da vari siti come quello del Comitato Unitario Permanente degli Ordini e Collegi Professionali [2], Confprofessioni [3] e Istat [4] .

[2]http://www.cuprofessioni.it
[3]https://confprofessioni.eu
[4]http://dati.istat.it

Per le parole relative ai lavori *ungendered* il calcolo dello *score* era semplicemente il risultato della proiezione del vettore lungo la direzione, mentre per le parole relative ai lavori *gender* sono state effettuate due analisi: la prima ha utilizzato la formula (3.2), nella quale si calcolano gli *score* di entrambe le versioni della parola e lo *score* finale è la media dei due *score* ottenuti; la seconda analisi, invece, ha tenuto conto degli *score* di ogni componente in maniera separata. Le figure Figura 4.1 e Figura 4.2 mostrano quanto appena descritto proiettando i WEs lungo la direzione lui-lei.

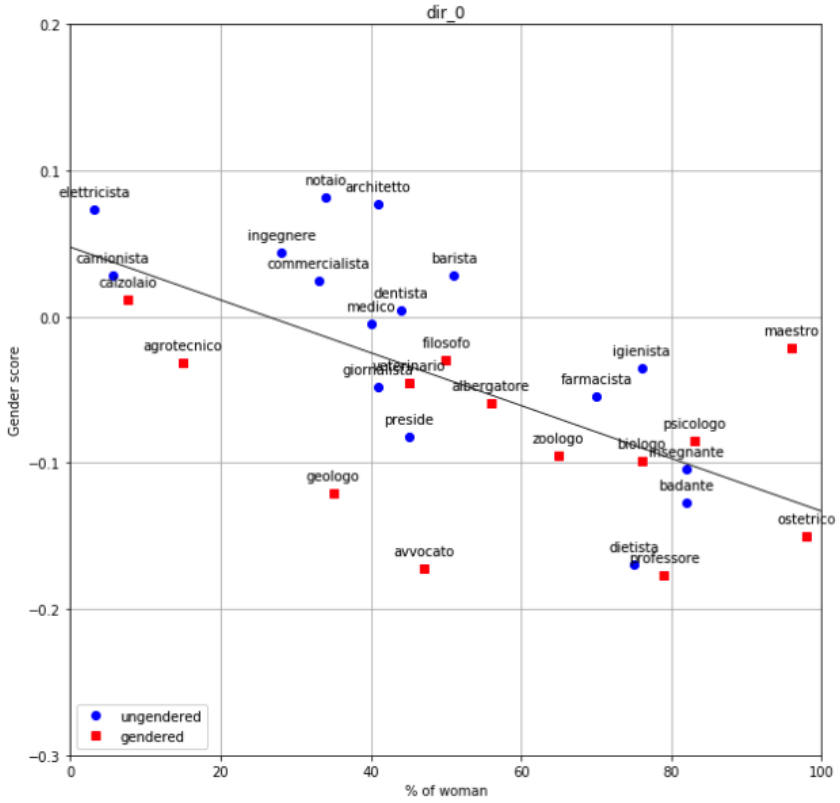

Figura 4.1: Ungendered works projected on direction lui-lei

Per completezza si riportano anche i grafici ottenuti proiettando le parole relative alle occupazioni lungo la direzione ottenuta come sommatoria delle prime due componenti risultanti dall'analisi PCA.

Questo primo passo di analisi del funzionamento dei WEs è stato propedeutico per quanto segue. L'analisi del gender *score* dei WEs è stata estesa, infatti, ad un documento completo.

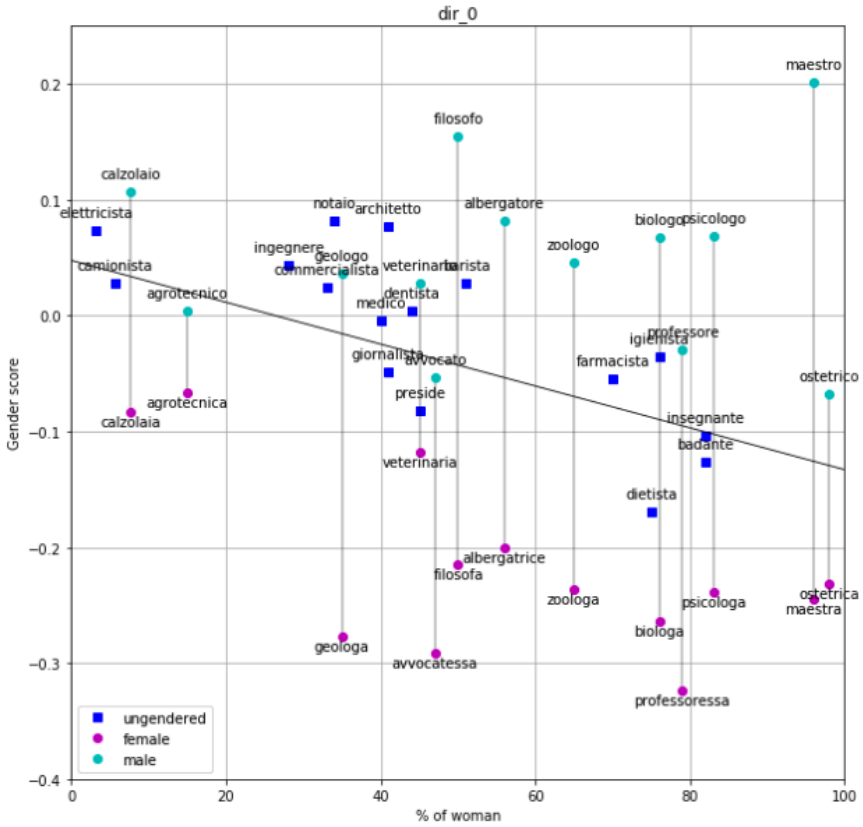

Figura 4.2: Gendered works projected on direction lui-lei

Tale analisi è partita con la rimozione dei caratteri di interpunzione e la determinazione di una lista di *stopwords*, ovvero parole come articoli, pronomi, congiunzioni, avverbi, verbi che risultavano essere rumore all'interno del documento. Mediante il *parsing* del documento, ogni parola appartenente alla lista di *stopwords* è stata rimossa dal documento originale. Alle parole rimanenti è stato applicato il modello definito in precedenza ed è stato calcolato il relativo WEs. In seguito, è stato proiettato il vettore ottenuto lungo la direzione lui-lei ed è stato calcolato lo *score* risultante. La media di tutti gli *score* delle parole nel documento rappresentava il *gender score* finale del testo.

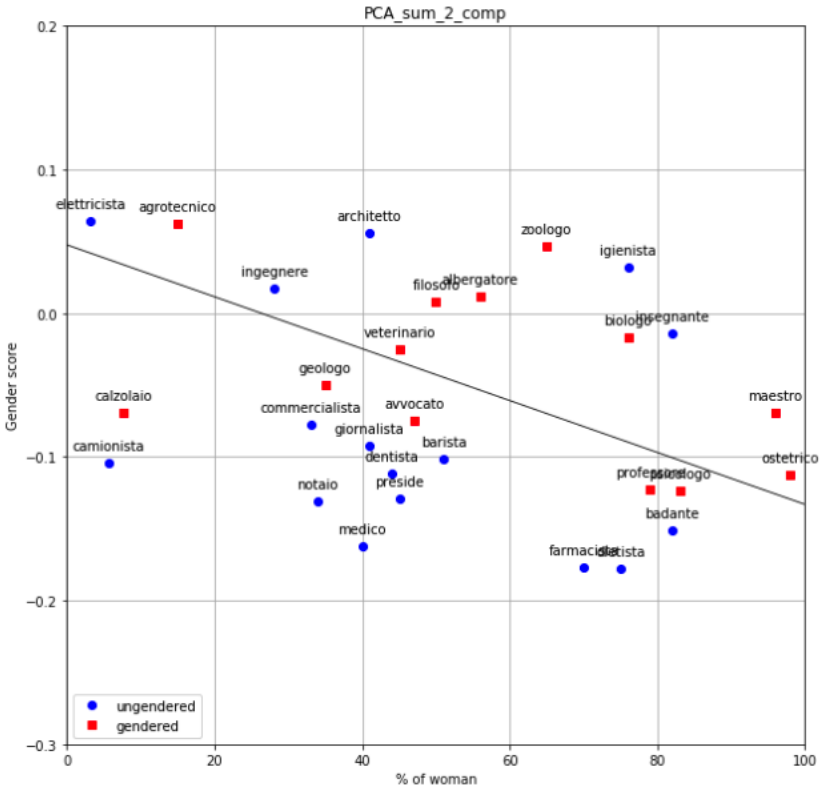

Figura 4.3: Ungendered works projected on direction sum PCA

I documenti analizzati con questa tecnica sono stati trentotto e trattavano di automobili, moda, formazione, *recruiting*, persone, relazioni, responsabilità e territorio. Questi ultimi cinque temi in particolare fanno riferimento a testi presenti nel sito di Miriade s.r.l [5] e descrivono alcuni fondamenti etici che l'azienda si prefissa di promuovere e mantenere con i propri clienti e con i propri dipendenti.

[5] https://www.miriade.it

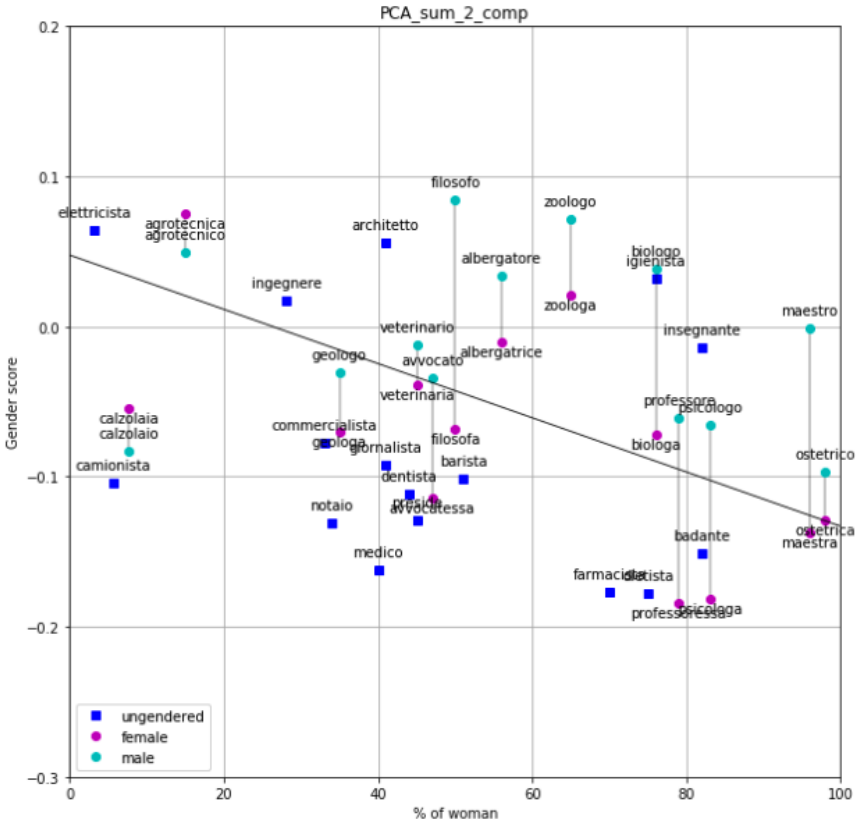

Figura 4.4: Gendered works projected on direction sum PCA

4.2 N-grams analysis

Nell'ambito del *text mining* e dell'NLP, gli n-grams sono sequenze contigue di N elementi presi da un testo o discorso parlato. Tipicamente queste sequenze si muovono di una parola alla volta. Per spiegare meglio questo fatto vediamo il seguente esempio:

Frase: "La macchina procede lungo la strada"

Dimensione n-grams = N = 2

Gli n-grams generati saranno i seguenti:

- la macchina;
- macchina procede;
- procede lungo;
- lungo la;

- la strada.

In questo caso abbiamo 5 n-grams. Come si può vedere, ci si sposta di una parola alla volta dall'inizio della frase verso la sua conclusione. Se, ad esempio, avessimo scelto N=3 come dimensione degli n-grams, il risultato sarebbe stato:

- la macchina procede;
- macchina procede lungo;
- procede lungo la;
- lungo la strada.

E si sarebbe continuato in tal modo per 4-grams, ecc. Nel caso in cui N=1 l'elemento viene chiamato *unigrams*, per N=2 viene definito *bigrams*, per N=3 *trigrams*, ecc.

Per determinare il numero di n-grams presenti in una frase con X parole e data una certa dimensione N basta utilizzare la seguente formula:

$$Ngrams_K = X - (N - 1) \tag{4.1}$$

Gli usi degli n-grams sono molteplici, ad esempio la *spelling correction*, il *word breking* e la *text summarization*. Spesso sono anche utilizzati quando si vuole sviluppare un *language model* e si vuole generare non solo l'*unigram model*, ma anche *bigram* o *trigram model*.

Gli esperimenti svolti in questo elaborato hanno utilizzato il pacchetto *ngrams* della libreria NLTK (Natural Language Toolkit), che, data una frase e la dimensione dell'n-grams, permette di ottenere elementi pronti per una successiva analisi.
Nel dettaglio, dato un documento, è stato effettuato il *parsing* del testo per la determinazione delle *sentence*. Per ognuna delle *sentence* trovate sono stati analizzati i rispettivi n-grams di dimensioni: 2, 3, 4, 5 e 6. Ad ogni *sentence* è stato attribuito uno *score*, ottenuto dalla media degli *score* dei singoli n-grams componenti la frase.

Infine, lo *score* del documento è stato calcolato dalla media degli *score* delle *sentence*. Di seguito viene riportato un grafico raffigurante il risultato dell'analisi n-grams sul documento *T1.txt*:

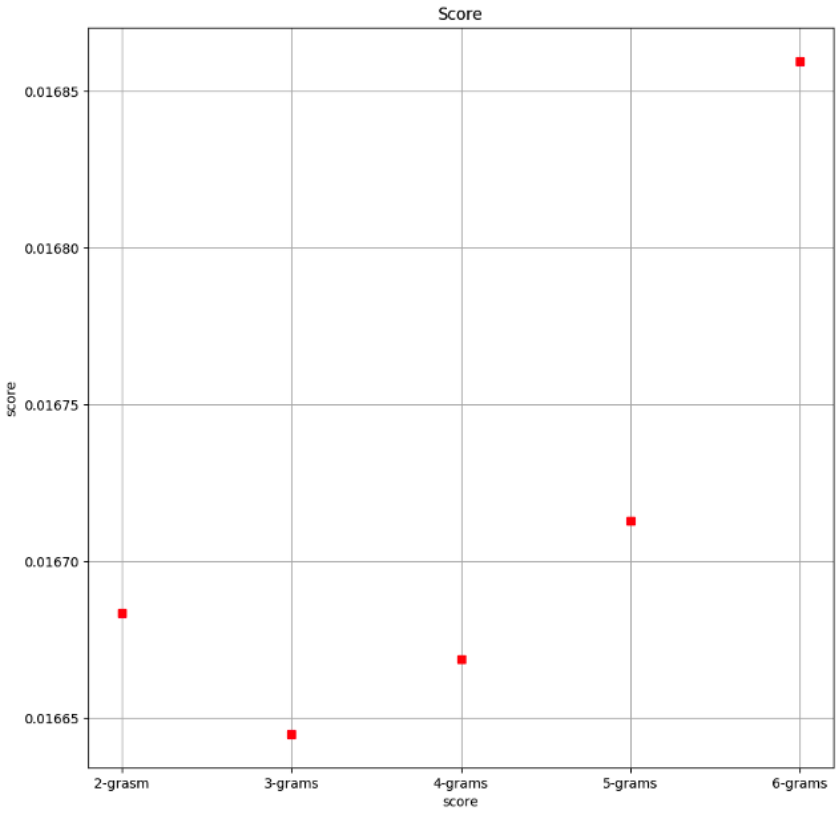

Figura 4.5: Analisi n-grams sul documento T1.txt

4.3 Sentence Embeddings analysis

I *Sentence Embeddings* possono essere visti come l'estensione del concetto che sta alla base dei WEs, ovvero rappresentare un'intera frase come un vettore numerico. Come i WEs, anche i *Sentence Embeddings* catturano relazioni semantiche tra frasi, come somiglianza, contraddizione o implicazione.

Ci sono vari modi tramite i quali si possono apprendere i *Sentence Embeddings* e uno di questi è l'utilizzo di algoritmi di ML, sia supervisionati, che non supervisionati. Questi algoritmi, risolvendo un certo compito di NLP utilizzando un set di dati etichettati (*supervised learning*), producono *Sentence Embeddings* universali

che possono essere a loro volta ottimizzati per diverse applicazioni. Generalmente è stato dimostrato che questi strumenti forniscono rappresentazioni semantiche molto più ricche di informazioni rispetto ai WEs.

Ad oggi sono diverse le tecniche di *Sentence Embeddings* che si possono utilizzare per ottenere una rappresentazione vettoriale della frase analizzata. Di seguito un breve elenco:

- *Universal Sentence Encoder (USE)*: un insieme di due modelli che, attraverso l'apprendimento multi-task, codifica le frasi in vettori di frasi altamente generici che risultano essere facilmente adattabili a molte attività di NLP;
- *SentenceBert:* basato sul modello BERT, combina la potenza delle architetture a trasformatori e delle *twin neural network* per creare rappresentazioni delle frasi di alta qualità;
- *SkipThought:* un adattamento di Word2Vec che produce *embeddings* imparando a prevedere i dintorni di una frase codificata;
- *InferSent:* produce *Sentence Embeddings* addestrando *neural network* al fine di identificare relazioni semantiche tra frasi in maniera supervisionata.

Per l'analisi dei *Sentence Embeddings* si è deciso di utilizzare InferSent. Presentato nel 2018 da Facebook AI Research, è, come descritto prima, un tecnica supervisionata per la costruzione dei *Sentence Embeddings*. La principale caratteristica di questo modello è essere allenato su dati prodotti con la tecnica del *Natural Language Inference*, più precisamente sul dataset SNLI (Stanford Natural Language Inference). Questo dataset consiste in 570 mila frasi generate dall'uomo in inglese, etichettate manualmente in una delle tre categorie: *entailment* (successione logica), *contradiction* (contraddizione), *neutral*.
Il modello ha due versioni, una che utilizza GloVe e una da poco implementata che utilizza FastText. Per coerenza con i dati sui vettori estratti si è scelto di procedere con il modello implementato con FastText. Ogni frase, nel documento oggetto di

analisi, viene passata al modello che determina, tra le altre cose, una percentuale di importanza della singola parola all'interno della frase. Al fine del calcolo dello score del documento, queste percentuali sono state moltiplicate per il valore dello *score* ottenuto proiettando il WEs della parola lungo la direzione scelta, in questo modo si ottiene uno *score* della *sentence* pesato. Lo *score* finale del documento sarà la media degli *score* delle singole frasi. In Figura 4.6 si possono vedere le percentuali di importanza delle singole parole all'interno di una frase.

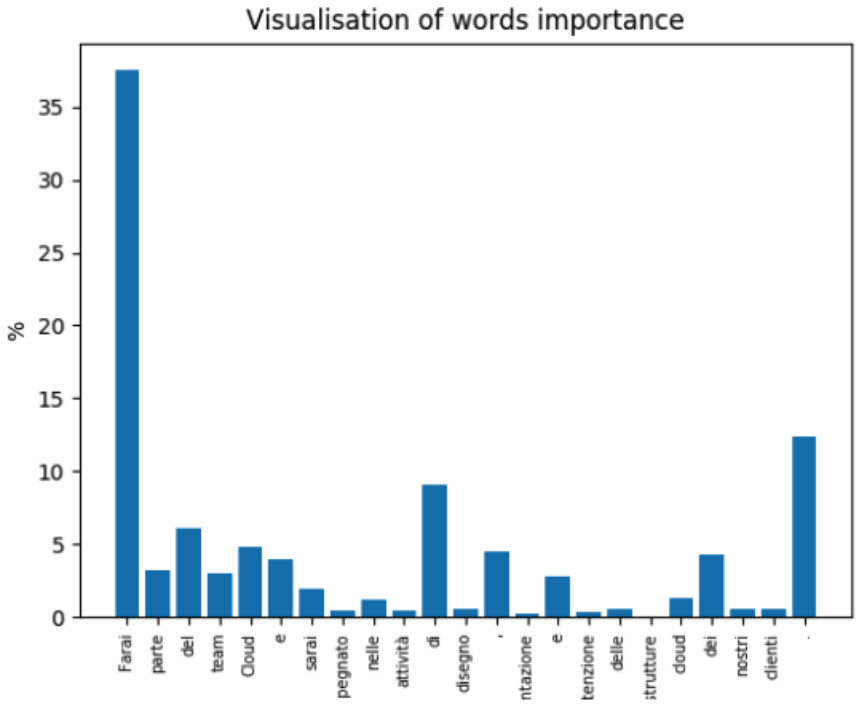

Figura 4.6: Analisi n-grams sul documento T1.txt

Capitolo 5

Analisi dei risultati

In questo capitolo verranno analizzati i risultati ottenuti dai vari test effettuati su trentotto differenti documenti testuali. Gli esperimenti sono stati eseguiti su un MacBook Pro (Retina, 13-inch, Early 2015) con processore 2,7 GHz Intel Core i5 dual-core, memoria 8 GB 1867 MHz DDR e scheda grafica Intel Iris Graphics 6100 1536 MB. Tutto il codice è stato scritto in linguaggio Python versione 3.8.1.

5.1 WEs results

Come descritto nel capitolo precedente, il primo esperimento effettuato è stato lo studio sul metodo di creazione e analisi dei WEs, ovvero il passaggio da parola a vettore. Un paio di premesse vanno fatte prima di procedere con l'analisi:

1. tutti i vettori sono stati normalizzati, di conseguenza è inevitabile aspettarsi risultati abbastanza vicini allo zero;

2. i modelli che mappano parole italiane in WEs non sono stati allenati, testati e studiati come i corrispondenti inglesi, di conseguenza un certo livello di errore è introdotto nei risultati.

Nella tabella seguente possiamo vedere un confronto tra lo *score* ottenuto dai vettori rappresentanti i lavori *ungedered* lungo la direzione lui-lei e la direzione ottenuta dalla somma vettoriale delle prime due componenti principali dell'analisi PCA.

Work	Lui-Lei	Sum Components
elettricista	0.07301048934459686	0.06400831176064473
camionista	0.02794816344976425	-0.10404465013750139
ingegnere	0.043688204139471054	0.017114373470745924
commercialista	0.024593153968453407	-0.07775996631108409
notaio	0.08197760581970215	-0.13119749473740966
architetto	0.07662126421928406	0.055500969355508016
dentista	0.004403832368552685	-0.111465794191953
medico	-0.004762651398777962	-0.16245962240972558
giornalista	-0.048219356685876846	-0.09273367557379224
barista	0.02783094346523285	-0.10171572098556357
igienista	-0.03521294146776199	0.03215335450829664
farmacista	-0.054419223219156265	-0.17667883372695142
preside	-0.08192481845617294	-0.12877952237558116
dietista	-0.1692095249891281	-0.17823189709710363
badante	-0.12698586285114288	-0.1515410373855104
insegnante	-0.10387009382247925	-0.014488489866103117

Tabella 5.1: Confronto direzioni su lavori ungendered

Come si può vedere, non tutti i risultati rispecchiano le attese, ovvero ottenere uno *score* quanto più vicino allo zero, questo, infatti, significherebbe che le parole sono *gender neutral*. Partendo dalla direzione lui-lei, le parole che meglio rappresentano un livello *gender neutral* sono dentista, con uno *score* gender male, e medico, con uno *score* negativo, quindi *gender female*. Vediamo, inoltre, una correlazione fra il *gender male* e le parole: elettricista, camionista, ingegnere, commercialista, notaio, architetto e barista, mentre le parole: giornalista, igienista, farmacista, preside, dietista, badante e insegnante tendono verso il *gender female*. Una possibile spiegazione di quanto ottenuto è che dentista e medico sono spesso rappresentati da entrambi i sessi, quindi nell'immaginario collettivo magari è più difficile identificare una netta distinzione fra il medico donna e il medico uomo e generalmente si fa riferimento ad entrambi i sessi con la stessa parola. Stesso discorso può essere fatto per la parola dentista. Per quanto riguarda, invece, le parole che identificano lavori più genderizzati, in entrambe le direzioni, sembra chiaro come questi risultati rispecchino la realtà del pensiero comune e di una parziale distribuzione reale dei lavoratori nei vari settori. Generalmente si fa riferimento ad un elettricista come uomo, nonostante i dati riportino che il 3,2% degli elettricisti sia donna, stessa cosa

accade anche nel caso del camionista, nonostante i dati segnino un 5,6% di donne che fanno questo lavoro, dato, tra l'altro, in crescita. Il medesimo ragionamento può essere fatto anche per lavori come l'insegnante, i dati, infatti, giustificano tale *bias*, dato che ben l'82% degli insegnanti è donna, nonostante l'aumento della percentuale di insegnanti uomini negli ultimi anni, o il/la badante (sempre l'82% di chi svolge questo lavoro è donna).

Passando alla direzione ottenuta dalla somma delle prime due componenti principali, si può vedere come i risultati siano profondamente diversi rispetto a quanto ottenuto dalla direzione lui-lei. In questo caso dodici *score* su sedici risultano essere *gender female* e solamente quattro *gender male*. I risultati, inoltre, presentano una forte genderizzazione in quasi tutti i casi.

Passiamo ora all'analisi degli *score* ottenuti dalle parole rappresentanti lavori con entrambe le declinazioni, sia maschile che femminile.

Work	**Lui-Lei**	**Sum Components**
calzolaio	0.011532962322235107	-0.06892190715133806
agrotecnico	-0.03136207535862923	0.0625957645218973
geologo	-0.1207280158996582	-0.05038159444668256
avvocato	-0.1724967509508133	-0.07466171515063812
albergatore	-0.059289515018463135	0.0116357783901709
veterinario	-0.04494456201791763	-0.02545074709690133
filosofo	-0.029997356235980988	0.008045279363555376
zoologo	-0.09548187255859375	0.04633926307746038
biologo	-0.09846513718366623	-0.016948613144893192
professore	-0.176707461476326	-0.1225756548701343
psicologo	-0.08505240827798843	-0.12361338673435736
ostetrico	-0.1500142216682434	-0.11296060157934441
maestro	-0.021833159029483795	-0.06920425323080294

Tabella 5.2: Confronto direzioni su lavori gendered mean

I risultati riportati rappresentano un tentativo di riduzione del *gender bias*. Gli *score* presenti nella Tabella 5.2 sono stati calcolati utilizzando [3.2], mediando, quindi, gli *score* della declinazione maschile e femminile della parola. Questo procedimento ha cercato, dunque, di ridurre il *bias* di genere, presente nella parola, tenendo

conto di entrambe le componenti.

Il calcolo della media dei due *score* relativi alle declinazioni maschile e femminile di una parola dovrebbe, nella teoria, portare ad uno *score* molto vicino allo zero, eliminando il *bias* di genere. Definendo un range di valori nei quali si può ritenere adeguato quanto ottenuto, ad esempio tra -0.05 e +0.05, vediamo che questa tecnica funziona in maniera adeguata in cinque casi su tredici, con una correttezza del 40% circa.

I risultati sono in linea con quanto ci si potrebbe aspettare, tenuto conto del fatto che non sempre vengono restituiti vettori corretti dal modello, di conseguenza non sempre le due declinazioni della parola sono mappate correttamente e non sempre, anche se mappate correttamente, sono alla stessa distanza, in valore assoluto, dalla direzione lungo la quale vengono proiettati i rispettivi vettori.

Infine, in Tabella 5.3, è possibile vedere, in particolare lungo la direzione Lui-Lei, come le diverse declinazioni delle parole, maschile e femminile, siano correttamente mappate. La mappatura corretta delle parole corrisponde ad uno *score* negativo se la parola è *gender female*, ad uno *score* positivo se *gender male*. Questo avviene correttamente in dieci casi su tredici. Osservando invece la direzione *sum components*, vediamo come tale direzione rappresenti correttamente questa distinzione di genere in solamente tre casi su dieci.

Word	Gender	Lui-Lei	Sum Components
calzolaio	calzolaio	0.10632234811782837	-0.08351331598730682
	calzolaia	-0.08325642347335815	-0.05433049831536929
agrotecnico	agrotecnico	0.004385811742395163	0.049778419445163015
	agrotecnica	-0.06710996478796005	0.07541310959863158
geologo	geologo	0.0358298122882843	-0.030444143028869736
	geologa	-0.2772858440876007	-0.07031904586449539
avvocato	avvocato	-0.05364465340971947	-0.034791220138460796
	avvocatessa	-0.2913488447666168	-0.11453221016281544
albergatore	albergatore	0.08156101405620575	0.03368751291614961
	albergatrice	-0.2913488447666168	-0.01041595613580781
veterinario	veterinario	0.028080355376005173	-0.011863708792899
	veterinaria	-0.11796947568655014	-0.03903778540090366
filosofo	filosofo	0.15449324250221252	0.08413044816491984
	filosofa	-0.2144879549741745	-0.06803988943780909
zoologo	zoologo	0.04532330855727196	0.07168404758470806
	zoologa	-0.23628705739974976	0.020994478570212697
biologo	biologo	0.06694553792476654	0.03824886680148145
	biologa	-0.263875812292099	-0.07214609309126783
professore	professore	-0.029695630073547363	-0.06087001063440093
	professoressa	-0.3237192928791046	-0.18428129910586769
psicologo	psicologo	0.0684058666229248	-0.06538849019471915
	psicologa	-0.23851068317890167	-0.18183828327399557
ostetrico	ostetrico	-0.0683940052986145	-0.09691320969541914
	ostetrica	-0.2316344529390335	-0.1290079934632697
maestro	maestro	0.201321080327034	-0.0014001445870383395
	maestra	-0.2449873983860016	-0.13700836187456755

Tabella 5.3: Confronto direzioni su lavori gendered

5.2 Document analysis

Come descritto in precedenza, lo scopo di questo elaborato è quello di fornire un primo approccio all'analisi dei *gender bias* nei documenti di testo. L'approccio più intuitivo e basilare è definire una direzione di genere e su questa proiettare i vettori delle parole in esso contenute. Di seguito vengono riportati i vari passaggi eseguiti per l'analisi del documento *T1.txt*.

Il primo è consistito nell'eliminazione delle *stopwords*. Il documento analizzato presentava 595 parole, mentre le parole effettivamente analizzate dopo aver processato il documento erano 285, 196 dopo la rimozione dei duplicati, quindi circa il

32%. Lo *score* medio del documento era 0.009647138 così distribuito:

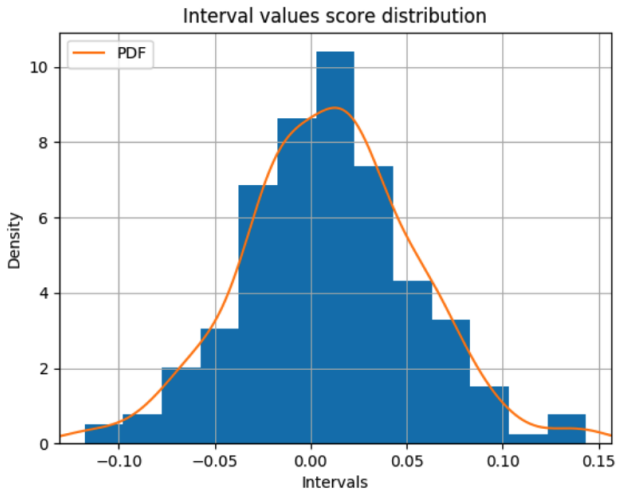

Figura 5.1: Distribution score document T1.txt

Lo *score* ottenuto non rappresenta correttamente il livello *gender* del documento, che è fortemente *gender male*. Per cercare di estrarre questa informazione, guardando alla distribuzione degli *score*, è stato determinato un *range* entro il quale gli *score* erano ritenuti neutri. Per il documento analizzato tale *range* è stato definito dall'intervallo -0.03 e 0.03. I valori considerati sono stati, quindi, solamente quelli al di fuori di tale *range*, pertanto il nuovo *score* è stato calcolato come la media di 84 valori, risultando pari a 0.02202492. Confrontato con il precedente valore ottenuto, questo *score* rispecchia in maniera più appropriata il reale *gender level* del documento.

Un'altra analisi, effettuata sempre sugli *score*, ha considerato, di volta in volta, una percentuale crescente degli *score* maggiormente genderizzati, in valore assoluto, per poi calcolare lo *score* solamente della percentuale di valori considerata. Questo ha consentito di analizzare come, all'aumentare del numero di *score* considerati, e di conseguenza di parole meno genderizzate, lo *score* del documento tendesse verso una direzione gender o verso la neutralità.
Come si può vedere dalla Tabella 5.4, al crescere dei valori considerati il documento tende ad un *gendel level neutral*, sintomo che gran parte dei termini mappati dal modello risultano avere uno *score* molto vicino allo zero.

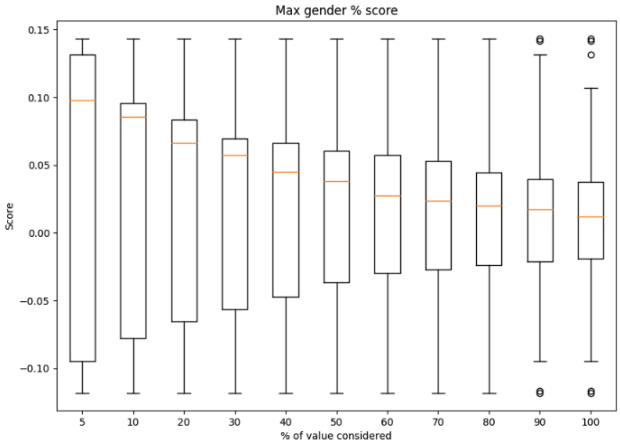

Figura 5.2: Distribution score document T1.txt abs analisys

Percentuale	Valori considerati	Score
5	9	0.042822465
10	19	0.030353678
20	39	0.031611204
30	58	0.024052545
40	78	0.022777375
50	98	0.017660942
60	117	0.015419679
70	137	0.014028817
80	156	0.012084265
90	176	0.010576802
100	196	0.009647138

Tabella 5.4: Risultati analisi percentuale

Lo step successivo proposto per l'analisi del *gender level* di un documento è consistito nell'analisi degli n-grams. Per effettuare quest'analisi è stato utilizzato il pacchetto *ngrams* della libreria *NLTK*.

Dopo aver suddiviso il documento in frasi (ogni frase è delimitata da un carattere "a capo"), ognuna di queste è stata ripartita in n-grams. Lo *score* della frase è stato calcolato dalla media degli *score* degli n-grams che la componevano. Infine, lo *score* del documento è stato dato dalla media degli *score* delle singole frasi. In particolare, il documento *T1.txt* è risultato avere uno *score* di 0.01668345, considerando n-grams di dimensione 2. La stessa procedura è stata effettuata anche con n-grams di dimensione 3, 4, 5 e 6.

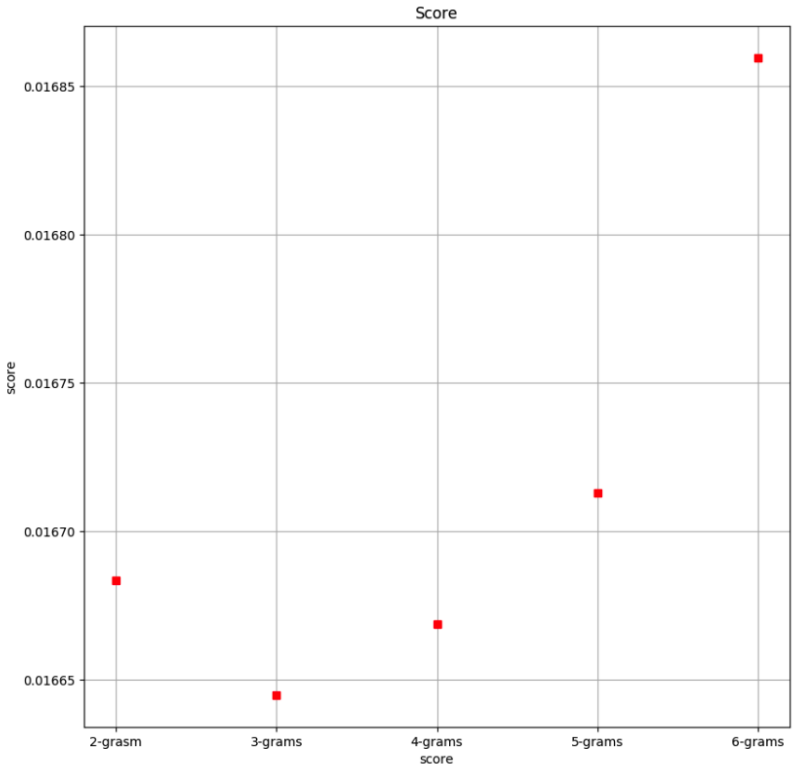

Figura 5.3: N-grams score document T1.txt

Come si può vedere in Figura 5.3, per n-grams di dimensione 3 c'è una tendenza *gender neutral* rispetto a n-grams di dimensione 2, 4, 5 e 6.

L'ultima analisi del documento per la determinazione del suo *gender level* si è basata sui *Sentence Embeddings*. Come descritto nei capitoli precedenti, il modello utilizzato è stato InferSent, in particolare InferSent2, che sfrutta la libreria FastText.

I risultati ottenuti sono stati due: il primo considera tutte le percentuali di importanza restituite dal modello InferSent, il secondo, invece, considera solo le percentuali di importanza inferiori all'1%, questo perché è stato riscontrato che la maggior parte di parole che effettivamente rappresentavano il genere del documento avevano una percentuale sotto questa soglia.

Il primo valore ottenuto è 0.05346934156609007, mentre il secondo è pari a 0.012834097075457865.

Quanto ottenuto dimostra come nel primo caso, quello in cui tutte le percentuali restituite dal modello vengono considerate, ci sia un forte aumento del *gender le-*

vel del documento rispetto a quanto emerso dalle precedenti analisi (score di circa 0.02), mentre nel secondo caso il risultato sia più coerente con i risultati trovati in precedenza.

5.2.1 Confronto risultati

Con questa tecnica sono stati analizzati in tutto trentotto documenti di testo così suddivisi:

- undici gender female;
- undici gender male;
- undici gender neutral;
- cinque riguardanti il sito web di un'azienda.

Di seguito viene riportata una tabella con i risultati ottenuti dall'analisi dei *Sentence Embeddings*:

Genere	**Documento**	**Score tot %**	**Score % < 1%**
Femminile	collant	0.057135653840842425	0.012832960693550265
	ferragni	0.026487593755042223	0.007887273602833462
	orecchini	0.05840921468164508	0.01480983090232998
	sandali	-0.003030775081747973	0.011149406628306782
	storiamoda	0.09504220465114518	0.007623342896805577
	verde	0.018613459538153406	0.007362597852938288
	sessanta	0.043913819817979684	0.013769686850103177
	lana	-0.009824361689064852	0.0030250755504048405
	capelli	0.07156032936345151	0.014458512062104668
	crema	0.13696633432758684	0.011342024026288283
	T2	0.04361938213038234	0.010396103683853936
Maschile	iwcorologio	0.09149366220337977	0.014083092511209214
	motogp	0.09261932959543054	0.024343042006172134
	movesion	0.04805116063162658	0.01226850349456278
	purosangue	0.06793212348125997	0.01977804627965622
	nissan	0.09142156934186306	0.021839523847055196
	kia	0.06356484822706426	0.01699635843814683
	enjoy	0.056523344751997676	0.018067383895756784
	sondaggio	0.08273959915360493	0.019569939679681087
	motore	0.06385659336335298	0.01748917547758791
	T1	0.05346934156609007	0.012834097075457865
Neutro	cervello	0.07072161632822606	0.017239819040040263
	geografia	0.04084453869765462	0.012425935457792086
	inclusione	0.004081035439733725	0.006717777659260619
	mangiamo	0.05804951421658223	0.010802143561088016
	ucraina	0.050463934624843915	0.013230244877888441
	citta	0.032062210149419984	0.0110719527312517
	dna	0.08992605512402417	0.01586471434428059
	unicore	0.07076275572163108	0.012355960133543786
	diabete	0.027284026644619982	0.0045776734607586335
	risparmio	0.11022031436353501	0.014406002979677346
	T3	0.04162215072544698	0.015913109164745934

Tabella 5.5: Analisi risultati documenti

Dalla tabella sopra riportata è possibile notare come ci sia una sostanziale differenza tra i risultati ottenuti considerando tutte le percentuali di importanza restituite dal modello e quelli riportati tenendo conto solamente delle percentuali inferiori all'1%. Analizzando ogni genere singolarmente, notiamo che, per quanto riguarda il genere femminile, solamente due casi su undici, e solo nel caso di tutte le percentuali

considerate, vengono mappati con segno negativo. Tale risultato può dipendere da diversi fattori, il primo che risulta interessante è l'assegnazione di importanza alle parole da parte di InferSent, dove, nella maggior parte dei casi, alte percentuali venivano assegnate a parole come il, un, uno, dei, che portano poca informazione al livello gender del documento o a parole *gender male*. Secondo fattore, sicuramente influente, è come il modello di FastText associ i vettori alle parole, mappando, ad esempio, parole *gender female* in vettori che poi risultano avere uno *score* positivo.

Passando ai documenti maschili, vediamo come ci sia una forte attenuazione del livello gender con la seconda modalità di analisi. Certamente i risultati ottenuti, mantenendo tutte le percentuali di importanza, rafforzano il presupposto che i documenti siano *gender male*, ma è altrettanto vero che ci si aspetta che venga effettuato un controllo sui documenti da parte degli editori sul grado di inclusione del documento stesso. In questa seconda prospettiva, tenendo come range di neutralità gli *score* che vanno da -0.015 a + 0.015, si può dedurre che i testi analizzati siano comunque fortemente genderizzati, risultano, infatti, neutri solamente tre documenti su undici.

Osservando, infine, i documenti neutri, dove tutti i testi sono stati presi da articoli presenti nel sito dell'Università di Padova, si può notare come ben otto documenti su undici risultino essere neutri, ovvero il 73%, risultato che può ritenersi in linea con i notevoli sforzi fatti dall'Ateneo negli ultimi anni per promuovere una divulgazione più inclusiva.

Nella Tabella 5.6 è stato riportato un confronto fra i risultati ottenuti su tutti i documenti con i tre strumenti utilizzati: WEs, n-grams e *Sentence Embeddings*. Definire uno strumento migliore di un altro risulta difficile, dato che sono tutte tecniche che, per quanto riguarda la lingua italiana, sono ancora in fase di studio e i modelli sui quali si basano sono principalmente allenati su testi in inglese. In ogni caso è abbastanza chiaro come i WEs, per quanto fatto in questa tesi, risultino lo strumento che meglio riesce ad individuare il livello *gender* di un documento.

Genere	Documento	WEs	2-grams	Sentence Embeddings
Femminile	collant	-0.014407867	0.0057087783	0.012832960693550265
	ferragni	-0.004810314	0.0054914346	0.007887273602833462
	orecchini	-0.009161278	0.018470164	0.01480983090232998
	sandali	-0.010615156	0.006429503	0.011149406628306782
	storiamoda	-0.0006078693	0.008214074	0.007623342896805577
	verde	-0.0028842464	0.015779084	0.007362597852938288
	sessanta	0.011482517	0.02003216	0.013769686850103177
	lana	-0.009865139	0.008507842	0.0030250755504048405
	capelli	0.007842251	0.018303085	0.014458512062104668
	crema	-0.00956937	0.008841601	0.011342024026288283
	T2	-0.011754412	0.014438667	0.010396103683853936
Maschile	iwcorologio	0.0125394575	0.017765088	0.014083092511209214
	motogp	0.03067494	0.032462187	0.024343042006172134
	movesion	0.021565478	0.018681861	0.01226850349456278
	purosangue	0.019303998	0.019731972	0.01977804627965622
	nissan	0.029163295	0.027232695	0.021839523847055196
	kia	0.008914242	0.018882511	0.01699635843814683
	enjoy	0.022353152	0.02461968	0.018067383895756784
	sondaggio	0.020924723	0.023403304	0.019569939679681087
	motore	0.033683658	0.027340189	0.01748917547758791
	T1	0.02202492	0.01668345	0.012834097075457865
Neutro	cervello	0.014221825	0.0197271	0.017239819040040263
	geografia	0.010693957	0.015065374	0.012425935457792086
	inclusione	0.011967082	0.018971564	0.006717777659260619
	mangiamo	0.0025916363	0.013395935	0.010802143561088016
	ucraina	0.020322533	0.022404682	0.013230244877888441
	citta	0.00087702484	0.011344261	0.0110719527312517
	dna	0.011420355	0.014694984	0.01586471434428059
	unicore	0.007308646	0.0139123	0.012355960133543786
	diabete	0.002244793	0.009792053	0.0045776734607586335
	risparmio	-0.0008712055	0.01776114	0.014406002979677346
	T3	0.009495191	0.019136332	0.015913109164745934

Tabella 5.6: Confronto strumenti

Si riferiscono alcune percentuali ritenute significative: per i documenti *gender female* i WEs mappano correttamente nove documenti su undici, cioè l'82%, per i testi *gender male* i risultati esatti sono otto su undici, tenendo conto del range di neutralità definito in precedenza, ovvero il 73%, mentre per i *gender neutral* sono ben dieci su undici i documenti corretti, quindi il 90%.

La colonna relativa ai *Sentence Embeddings* è già stata trattata in precedenza, in

ogni caso per quanto riguarda i documenti maschili e neutri i risultati paiono essere decisamente attendibili.

Infine, una considerazione per quanto riguarda gli n-grams. Questo è uno strumento sicuramente utile, che in alcuni casi viene utilizzato per allenare le reti che poi restituiscono informazioni per i *Sentence Embeddings*, come nel caso di InferSent. Probabilmente, il fatto di andare a confrontare le parole vicine non sempre aiuta ad identificare il genere, soprattutto quello femminile, ma probabilmente è un ottimo metodo per studiare, ad esempio, la similarità fra parole vicine, di conseguenza quanto le parole in una frase siano inerenti, ad esempio, allo stesso contesto di cui si sta parlando. Nella Tabella 5.6 sono stati riportati i risultati riguardanti n-grams con $n = 2$, ma l'analisi è stata più ampia considerando anche n-grams di dimensione 3, 4, 5, 6, come detto in precedenza. I risultati si possono suddividere in due categorie: una contempla un miglioramento dello score del documento all'aumentare della dimensione dei *grams*, come si può vedere in Figura 5.5, l'altra, aumentando la grandezza dei *grams*, peggiorava, come in Figura 5.4.

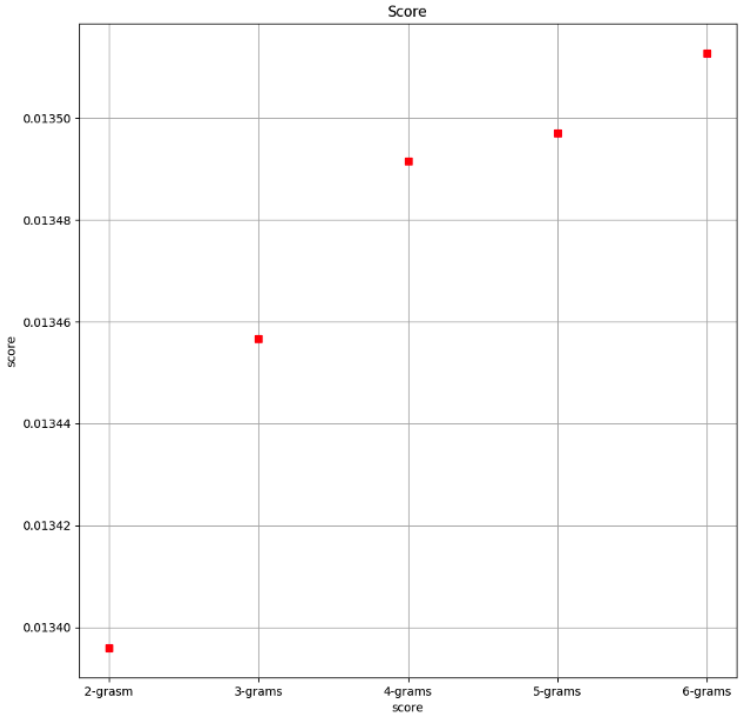

Figura 5.4: N-grams mangiamo.txt

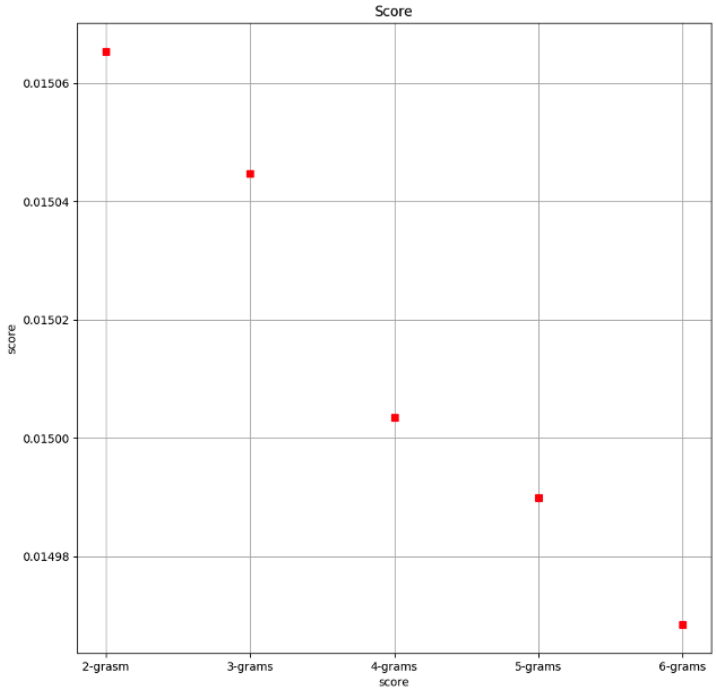

Figura 5.5: N-grams geografia

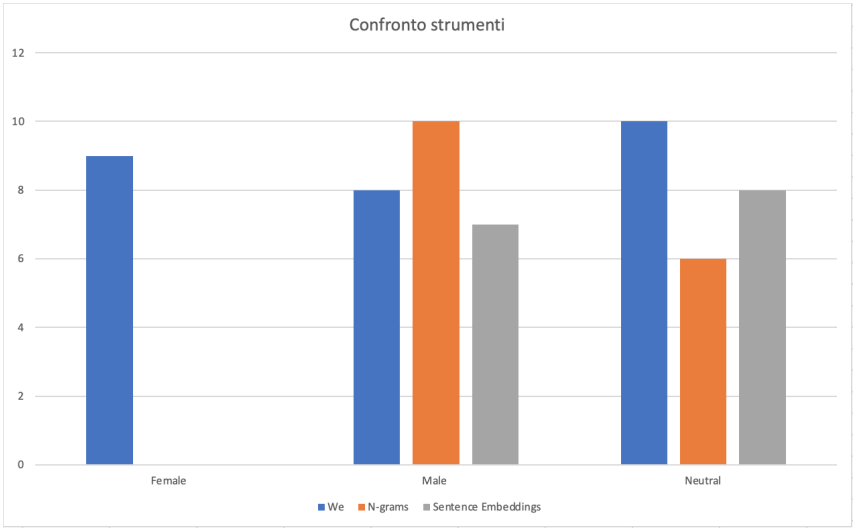

Figura 5.6: Confronto correttezza strumenti

5.3 Miriade: un caso di studio

Dal 7 Marzo 2022 ho cominciato uno stage curriculare presso Miriade s.r.l., azienda di consulenza informatica con sede a Thiene (VI). Miriade non è la solita azienda di consulenti 9-18, bensì un gruppo di informatici, matematici, statistici, ingegneri, creativi e filosofi che uniscono le proprie competenze per trovare la soluzione più adatta alle esigenze di ogni cliente. Play, Passione, Unconventional, Integrità e Disponibilità sono i cinque valori che rappresentano questa società e tutti i suoi dipendenti. I servizi offerti spaziano dal Data Management (DBA & DataOps, Data Integration e Data Governance), al Digital Enterprise (Cloud, DevOps e Modern App) e alla Smart Analitycs (Business Intelligence, Process automation e Data Science & ML).
Miriade non è attenta solamente ai propri clienti, presta particolare attenzione al proprio territorio sostenendo le Botteghe del Mondo e associazioni nazionali come il CIAI, Centro Italiano Aiuti all'Infanzia, e alle persone che lavorano all'interno dei vari team. Proprio da questa particolare attenzione è nato il Codice Etico Informale, che descrive e riassume i valori di Miriade e le modalità attraverso le quali si intende applicarli.

Assieme ad alcune colleghe è nata l'idea di far diventare Miriade un caso di studio all'interno di questo elaborato, per cercare di determinare quanto il sito internet dell'azienda presenti testi *gender neutral* e di ridurre, o meglio ancora eliminare, eventuali *bias* di genere presenti.

L'analisi si è soffermata su cinque testi presenti all'interno del sito:

- un'offerta di lavoro per un posto come Cloud Engineer;
- miriade e le persone che compongono i vari team;
- l'attenzione per il territorio;
- le relazioni con i clienti;
- la responsabilità di lasciare un segno positivo in quello che si fa.

I testi sono stati estratti dal sito e salvati in documenti testuali per essere poi elaborati dallo strumento di analisi costruito. Vengono riportati i risultati ottenuti dall'analisi WEs e l'analisi dei *Sentence Embeddings* considerando solamente le percentuali minori all'1%, che risultavano essere i più appropriati dopo un confronto con quanto ottenuto anche dalle analisi degli altri documenti studiati:

Documento	**Score WEs**	**Score Sentence**
Cloud Engineer	0.03497703	0.026026363251629634
People	0.016961666	0.015696057148957592
Relazioni	0.010022098	0.016215468852241618
Responsability	0.015484632	0.020263223822801043
Territorio	0.0024016262	0.010991989314330344

Tabella 5.7: Risultati analisi testi Miriade

Definendo un intervallo da $\sim$ (-0.015) e $\sim$ (0.015), intervallo che può essere considerato *gender neutral*, dopo aver analizzato i documenti e i relativi *score* attribuiti alle parole, vediamo come dall'analisi WEs ben tre documenti su cinque risultino *gender neutral*, mentre per i *Sentence Embeddings* il risultato si riduca leggermente a due su cinque (anche se possono essere considerati tre su cinque con uno scarto dello 0.01). Questi risultati dimostrano come l'azienda abbia lavorato, e continui a lavorare, per fare dell'inclusione e del rispetto dell'identità delle persone che vengono a contatto con la realtà di Miriade, uno dei propri valori.

Conclusioni

In quest'ultimo capitolo verrà riassunto quanto fatto, riportando vantaggi e svantaggi degli strumenti utilizzati e discutendo di possibili lavori futuri legati alla tematica presentata.

Sintesi

Nel campo dell'NLP, *bias* di genere e stereotipi sono argomenti che hanno da poco preso piede, di conseguenza gli studi sono ancora in fase iniziale e si ricercano continuamente nuove modalità per la determinazione di questi problemi, motivo per il quale l'applicazione dei modelli maggiormente studiati e comprovati porta con sé la possibilità di amplificare questi errori.

All'interno di tale lavoro, è stato dimostrato come sia possibile, attraverso l'utilizzo della *cosine similarity* e la definizione di una direzione di genere, determinare se una parola sia *gender female*, *gender male* o *gender neutral*. In questo modo, dato un documento è possibile avere una stima del relativo livello di genere. Questo è stato identificato come passaggio iniziale per una prima analisi di un testo.

Il primo *upgrade* a questa metodologia sono gli n-grams, ovvero parole contigue in piccoli gruppi n-dimensionali, attraverso le quali viene calcolato lo *score* del documento. Tale tecnica, per i risultati ottenuti, non si è rivelata particolarmente efficace e i motivi sono stati precedentemente discussi. Gli n-grams, tuttavia, rimangono uno strumento decisamente valido e sfruttato soprattutto in fase di *training* dei modelli che poi vengono utilizzati per ritornare i WEs.

Infine, è stata analizzata la tecnica di determinazione del *gender level* attraverso i *Sentence Embeddings*. A livello teorico questa tecnica è quella che meglio dovrebbe riuscire a dedurre ed eventualmente mitigare eventuali *bias* di genere presenti in un documento, questo perché determina, per ogni frase (*sentence*), una percentuale di importanza, mitigando (o amplificando) lo *score* della singola parola. Il modello InferSent2 è stato adottato come strumento per questa analisi, dato che è stato allenato con vettori derivanti da FastText, garantendo così un'uniformità all'interno del lavoro di tesi.

Limitazioni

Allo stato attuale le limitazioni allo studio dei WEs italiani sono diverse.

La principale è forse la lingua italiana stessa, una lingua fortemente genderizzata, ricca di modi di dire e di molte terminologie derivate dai diversi dialetti. Questo sicuramente non aiuta la determinazione di eventuali *bias* di genere, dato che molto spesso una parola è ambigenere e viene identificata dall'articolo che la precedere, ad esempio *il* pediatra o *la* pediatra.

Altra limitazione sono i modelli stessi, nella maggior parte dei casi allenati su testi in inglese e poi riportati per parole in italiano, e caratterizzati da una scarsa precisione nel mappare le parole rispetto ad una direzione di genere identificata. Parole maschili mappate come femminili e viceversa, oppure parole neutre mappate come femminili o maschili. Un esempio è la mappatura, da parte del modello utilizzato per i WEs *cc.it.300*, della parola avvocato come femminile, o la parola professore.

Lavori futuri

La ricerca in questo campo continua, non solo per trovare nuovi modelli, ma anche, e soprattutto, per cercare di definire tecniche e strumenti per effettuare il *debiasing* dei WEs, degli n-grams, dei *Sentence Embeddings* (Kaneko, Bollegala, 2021; Liang, Li, Zheng, Lim, Salakhutdinov, Morency, 2020; Sun, Gaut, Tang, Huang, ElSherief, Zhao, Mirza, Belding, Chang, Wang, 2019; Bojanowski, Grave, Joulin, Joulin, 2016; Rathore, Dev, Srikumar, Phillips, Zheng, Yeh, Wang, Zhang, Wang, 2022; Fabris, Purpura, Silvello, Susto, 2020).

Il *debiasing* è sicuramente un campo dove la ricerca farà grandi passi in avanti, grazie alle nuove tecniche di ML e Deep Learning, al progresso delle Intelligenze Artificiali e ai nuovi strumenti hardware a disposizione dei ricercatori. Un'idea per quanto riguarda il *debiasing* della lingua italiana potrebbe essere quella di definire una sorta di dizionario, all'interno del quale vengono inserite tutte quelle parole che hanno sia la declinazione maschile che quella femminile, ad esempio dottore e dottoressa, e, nel momento in cui si analizza la parola dottore o dottoressa, all'interno di un testo, andare a calcolare non lo *score* relativo alla singola parola, bensì la media delle due, ottenendo così un valore mitigato fra le due componenti di genere.

Altro interessante campo di ricerca sono sicuramente i WEs relativi alla lingua italiana. Sarebbe, infatti, utile cercare di definire modelli che riescano a determinare in maniera più precisa eventuali stereotipi o *bias* di genere nei testi, alla stessa maniera di come fanno i corrispettivi strumenti per la lingua inglese.

Il lavoro futuro più stimolante riguarda, però, senza dubbio, una maggior educazione all'inclusione, all'attenzione per l'altro, per aiutare le persone a capire che, anche se spesso si pensa di essere responsabili di quello che si dice e non di quello che capiscono gli altri, in realtà si è soprattutto responsabili di quello che gli altri capiscono quando viene espressa un'opinione, una critica o un giudizio. Trovo significativo lo sforzo portato avanti dall'Università di Padova nel cercare di formare i propri studenti, facendoli riflettere e crescere anche relativamente a questo aspetto.

In particolare, nel mio corso di laurea, questo è stato possibile grazie all'impegno della professoressa Badaloni e dal professor Rodà e al loro corso di Saperi di Genere ed Etica nell'Intelligenza Artificiale, che invito, chiunque ne abbia la possibilità, a seguire. Tutto quello che, infatti, le intelligenze artificiali imparano, lo acquisiscono dalle persone che le progettano e le costruiscono. Siamo, dunque, ancora noi a fare la differenza, sempre.

Bibliografia

A. J. Cuddy, E. B. Wolf, P. Glick, S. Crotty, J. Chong, and M. I. Norton. (2015). "*Men as cultural ideals: Cultural values moderate gender stereotype content.*" In.

Adamuthe, Amol (ago. 2019). "Comparative Study of Convolutional Neural Network with Word Embedding Technique for Text Classification". In: *International Journal of Intelligent Systems and Applications* 11, pp. 56–67. DOI: 10.5815/ijisa.2019.08.06.

Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, Gian Antonio Susto (2020). "*Gender Stereotype Reinforcement: Measuring the Gender Bias Conveyed by Ranking Algorithms*". In.

Aliverti, Emanuele et al. (2021). "Removing the influence of group variables in high-dimensional predictive modelling". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184.3, pp. 791–811. DOI: https://doi.org/10.1111/rssa.12613. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12613.

Basta, Christine, Marta R. Costa-jussà e Noe Casas (apr. 2021). "Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings". In: *Neural Comput. Appl.* 33.8, pp. 3371–3384. ISSN: 0941-0643. DOI: 10.1007/s00521-020-05211-z. URL: https://doi.org/10.1007/s00521-020-05211-z.

Ben Schmidt (2015). "*Rejecting the gender binary: a vector-space operation*". In.

Berardi, Giacomo, Andrea Esuli e Diego Marcheggiani (2015). "Word Embeddings Go to Italy: A Comparison of Models and Training Datasets". In: *Italian Information Retrieval Workshop*.

Bojanowski, Piotr et al. (2016). *Enriching Word Vectors with Subword Information*. DOI: 10.48550/ARXIV.1607.04606. URL: https://arxiv.org/abs/1607.04606.

Bolukbasi, Tolga et al. (2016). *Quantifying and Reducing Stereotypes in Word Embeddings*. DOI: 10.48550/ARXIV.1606.06121. URL: https://arxiv.org/abs/1606.06121.

Boroditsky, Lera (2011). "How language shapes thought." In: *Scientific American* 304.2, pp. 62–65. ISSN: 1048-9843. DOI: https://doi.org/10.1016/j.leaqua.2003.09.004. URL: https://sites.unimi.it/zucchi/NuoviFile/Boroditsky-How%5C%20Language%5C%20Shapes%5C%20Thought.pdf.

Carsten Lygteskov Hansen, Melanie Tosik, Gerard Goossen, Chao Li, Lena Bayeva, Florence Berbain, Mihai Rotaru. (2015). "*How to Get the Best Word Vectors for Resume Parsing.*" In.

Chang, Kai-Wei, Vinodkumar Prabhakaran e Vicente Ordonez (nov. 2019). "Bias and Fairness in Natural Language Processing". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*. Hong Kong, China: Association for Computational Linguistics. URL: https://aclanthology.org/D19-2004.

Davide Biasion, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto (2020). "*Gender Bias in Italian Word Embeddings*". In.

Dev, Sunipa e Jeff Phillips (2019). *Attenuating Bias in Word Vectors*. DOI: 10.48550/ARXIV.1901.07656. URL: https://arxiv.org/abs/1901.07656.

Doughman, Jad e Wael Khreich (2022). *Gender Bias in Text: Labeled Datasets and Lexicons*. DOI: 10.48550/ARXIV.2201.08675. URL: https://arxiv.org/abs/2201.08675.

Doughman, Jad, Wael Khreich et al. (ago. 2021). "Gender Bias in Text: Origin, Taxonomy, and Implications". In: *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Online: Association for Computational Linguistics, pp. 34–44. DOI: 10.18653/v1/2021.gebnlp-1.5. URL: https://aclanthology.org/2021.gebnlp-1.5.

Eagly, Alice H e Linda L Carli (2003). "The female leadership advantage: An evaluation of the evidence". In: *The Leadership Quarterly* 14.6, pp. 807–834. ISSN: 1048-9843. DOI: https://doi.org/10.1016/j.leaqua.2003.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1048984303000584.

Eisenstein, Jacob et al. (nov. 2014). "Diffusion of Lexical Change in Social Media". In: *PLOS ONE* 9.11, pp. 1–13. DOI: 10.1371/journal.pone.0113114. URL: https://doi.org/10.1371/journal.pone.0113114.

Garg, Nikhil et al. (apr. 2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16. DOI: 10.1073/pnas.1720347115. URL: https://doi.org/10.1073%5C%2Fpnas.1720347115.

Giacomo Berardi, Andrea Esuli, and Diego Marchegiani. (2015). "*Word embeddings go to italy: A com- parison of models and training datasets.*" In.

J. T. Jost and A. C. Kay. (2005). "*Exposure to benevolent sexism and complementary gender stereotypes: consequences for specific and diffuse forms of system justification.*" In.

Kaneko, Masahiro e Danushka Bollegala (2021). *Debiasing Pre-trained Contextualised Embeddings*. DOI: 10.48550/ARXIV.2101.09523. URL: https://arxiv.org/abs/2101.09523.

Kollmayer, Marlene, Barbara Schober e Christiane Spiel (2018). "Gender stereotypes in education: Development, consequences, and interventions". In: *European Journal of Developmental Psychology* 15.4, pp. 361–377. DOI: 10.1080/17405629.2016.1193483. eprint: https://doi.org/10.1080/17405629.2016.1193483. URL: https://doi.org/10.1080/17405629.2016.1193483.

Liang, Paul Pu et al. (2020). *Towards Debiasing Sentence Representations*. DOI: 10.48550/ARXIV.2007.08100. URL: https://arxiv.org/abs/2007.08100.

Mikolov, Tomas, Wen-tau Yih e Geoffrey Zweig (giu. 2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Com-

putational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090.

Nissim, Malvina, Rik van Noord e Rob van der Goot (2019). *Fair is Better than Sensational:Man is to Doctor as Woman is to Doctor*. DOI: 10.48550/ARXIV.1905.09866. URL: https://arxiv.org/abs/1905.09866.

Rathore, Archit et al. (giu. 2022). "An Interactive Visual Demo of Bias Mitigation Techniques for Word Representations From a Geometric Perspective". In: *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. A cura di Douwe Kiela, Marco Ciccone e Barbara Caputo. Vol. 176. Proceedings of Machine Learning Research. PMLR, pp. 330–335. URL: https://proceedings.mlr.press/v176/rathore22a.html.

Rudkowsky, Elena et al. (2018). "More than Bags of Words: Sentiment Analysis with Word Embeddings". In: *Communication Methods and Measures* 12.2-3, pp. 140–157. DOI: 10.1080/19312458.2018.1455817. eprint: https://doi.org/10.1080/19312458.2018.1455817. URL: https://doi.org/10.1080/19312458.2018.1455817.

Sun, Tony et al. (2019). *Mitigating Gender Bias in Natural Language Processing: Literature Review*. DOI: 10.48550/ARXIV.1906.08976. URL: https://arxiv.org/abs/1906.08976.

T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013). "*Efficient estimation of word representations in vector space.*" In.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (2016). "*Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*". In.

Tripodi, Rocco e Stefano Li Pira (2017). *Analysis of Italian Word Embeddings*. DOI: 10.48550/ARXIV.1707.08783. URL: https://arxiv.org/abs/1707.08783.

Ringraziamenti

Alla fine di questo percorso credo sia giusto fermarsi e prendersi del tempo per ringraziare tutte quelle persone che in questi anni mi hanno accompagnato, supportato e sopportato.

Il primo ringraziamento lo devo a mamma Giusy e papà Carlo.

Grazie mamma per avermi insegnato cos'è la semplicità, il saper gioire delle piccole cose. In questi anni di università mi hai sempre ribadito di non prendermi troppi impegni perché dovevo studiare, c'era Camilla e dovevo dedicare tempo alle cose importanti. Grazie, perché con la tua presenza, mi ricordi quanto importante sia dare il giusto tempo alle persone che mi stanno accanto prima di pensare ai mille svaghi che la vita mi pone davanti.

Grazie papà perché in questi anni non mi hai mai fatto mancare il tuo appoggio, ricordandomi sempre quanto lo studio sia importante nella formazione di una persona. Grazie perché con il tuo esempio mi dimostri l'importanza di impegnarsi in quello che si fa. Sei e sarai sempre il mio punto di riferimento.

Un grazie particolare va a Matteo e Chiara, fratelli, amici, complici. Siete e sarete sempre i regali più belli che mamma e papà mi hanno fatto, nonostante tutto.

Grazie Camilla. Probabilmente se non fosse per la tua determinazione e il tuo continuo credere in me, oggi non avrei raggiunto questo traguardo. Grazie per avermi spinto ad intraprendere la carriera universitaria. Grazie per il supporto che in

ogni momento, a modo tuo mi dai. Grazie per essere la mia compagna di vita da 9 anni, siamo diventati grandi tenendoci per mano e chissà cosa ci riserverà il futuro, intanto ti ringrazio per tutto quello che fai per me.

Grazie ad Anass e Sara, compagni superstiti di questo viaggio. Senza di voi non sarebbe stato così speciale. Vi voglio bene.

Grazie alle mie due nonne, Sergia e Natalina. Presenze costanti nella mia vita. Sono certo che parte di questi risultati sono anche merito delle infinite preghiere che dite per me.

Un grande grazie va a tutti i miei amici, passati, presenti e anche a quelli che verranno. Vi ringrazio per tutti i momenti di festa che viviamo assieme, siete quel qualcosa in più nelle giornate grigie, e la ciliegina sulla torta in qualsiasi altro momento! Grazie!

Grazie a Maddalena e Massimo, mi avete accolto come un figlio in casa vostra e non perdete occasione per aiutarmi in qualsiasi cosa. Grazie.

Un grazie particolare va ad Arianna, collega e amica, che ha creduto in me in questi mesi di stage. Grazie Ari.

Un sentito ringraziamento va anche ad Alessandro Fabris, supporto indispensabile soprattutto nelle prime parti di questo lavoro. Grazie Alessandro e in bocca al lupo per i tuoi lavori presenti e futuri.

Infine devo ringraziare in particolar modo il Professor Rodà che in questi mesi di tesi è stato una guida fondamentale per la riuscita di questo lavoro.