Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea Magistrale in Informatica

Intelligenza Artificiale e Psicografia: come personalità e genere influiscono sulla scelta occupazionale

Tesi di laurea magistrale

Relatore
Prof. Nicolò Navarin

Laureanda
Eleonora Signor

Anno Accademico 2022-2023

"Rare sono le persone che usano la mente, poche coloro che usano il cuore e uniche coloro che usano entrambi."

— *Rita Levi Montalcini*

Dichiaro la sottoscritta e autore del presente documento come responsabile del suo contenuto, anche per le parti in esso riportate tratte da altre opere; queste ultime espresse citando le fonti.

Abstract

Il genere di appartenenza e la personalità possono influire sulla scelta occupazionale? In questa tesi si risponde a questa domanda utilizzando modelli di *Machine Learning* per l'analisi del linguaggio naturale su post pubblici di *Twitter*, per cogliere le relazioni esistenti tra genere, tratti di personalità *Big Five* e occupazione. Viene poi effettuato uno studio di fattibilità sull'utilizzo del *targeting* psicografico, strumento di comunicazione socio-demografico basato sui tratti di personalità, per influenzare le scelte di alcuni soggetti attraverso annunci pubblicitari sviluppati ad-hoc.

La tesi è applicata, nel concreto, nello studio delle occupazioni STEM e nell'individuare una soluzione che aiuti a risolverne il *gender gap*.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

In questo capitolo definiamo il problema affrontato da questa tesi e l'organizzazione del documento.

1.1 Definizione del problema

Numerosi studi compiuti soprattutto negli ultimi anni da Gruppi di ricerca e Università di tutto il mondo (si vedano, ad esempio, *Christianne Corbett e Catherine Hill 2015* [123]; *Commissione Europea 2019* [29]; *Fidelia Law et al. 2021* [80]; *Lina Aldén e Emma Neuman 2022* [11]; *Lucy Lu Wang et al. 2019* [152]; *Meyer M. et al. 2019* [88]; *Sarah-Jane Leslie et al. 2015* [84]) hanno evidenziato come vi siano differenze di genere rilevanti nella scelta occupazionale. In questo scenario appare fondata l'ipotesi che un campo prevalentemente maschile siano le carriere scientifiche, tecnologiche, ingegneristiche e matematiche (STEM) [29, 84, 88, 11] e nello specifico la disciplina Informatica [152]. È stato ipotizzato come sul divario di genere o *gender gap* incidano gli *stereotipi*[g]STEM [80, 84, 88] che associano la "brillantezza" nei campi STEM più agli uomini che alle donne, il che può minare la volontà di queste ultime nel perseguimento di tale carriera. Inoltre alcune analisi [123, 32] hanno messo in luce come le studentesse non sono attratte dall'Informatica in quanto non si percepiscono parte della comunità ingegneristica STEM. Per giunta risentono della mancanza di modelli di ruolo positivi e insegnanti che fungano da mentori.

Il *gender gap* è tuttavia una situazione che non deve essere generalizzata, in quanto di appartenenza alla sola epoca contemporanea e al mondo occidentale.
Il primo programmatore della storia, era in realtà una programmatrice: *Lady Ada Lovelace* [141]; una matematica inglese che alla fine del 1800 scrisse il primo programma per computer. Per i successivi 175 anni circa furono le donne a presentarsi come una risorsa attiva e pioniera in Informatica e Matematica; a prova di ciò nel 1960 il governo federale degli Stati Uniti stimò che la maggioranza dei programmatori del *software* in carriera erano di sesso femminile, con il 37% di donne del Paese impiegate nelle professioni matematiche e informatiche [141]. È durante questi anni che emerse la figura di *Mary Allen Wilkes* [141], informatica e avvocata statunitense che lavorò alla progettazione dell'attuale *personal computer*. Tuttavia a partire dal 1990 vi fu un'inversione di rotta [141, 84]; da prima con una perdita leggera di quota scesa al 35%, fino ad arrivare a oggi; in cui è stato stimato che solo il 26% delle donne è occupato in ambito STEM, di cui un 16% preferisce Biologia e Scienze delle vita e solo un 2%

Matematica e Informatica [32].
Nei Paesi di religione mussulmana sembra invece esserci un mutamento di tendenza rispetto all'occidente. In Oman, Arabia Saudita e Uzbekistan le donne con lauree scientifiche superano il 50% [154] arrivando al 70% in Iran [57] mentre l'Indonesia conta un 48% di ingegneri donne [154].

Uno strumento utile per aiutare a diminuire l'incidenza del *gender gap* nelle occupazioni può essere il *targeting* psicografico [91, 146]. Con *targeting* psicografico si intende la pubblicità mirata online che si basa, per la pubblicazione degli annunci, anche sui tratti di personalità del proprio pubblico potenziale. A tal fine abbiamo innanzitutto cercato di comprendere se esiste una qualche relazione tra genere, tratti di personalità e scelta occupazionale (*«Il genere di appartenenza e la personalità possono influire sulla scelta occupazionale?»*); per farlo abbiamo utilizzato l'analisi del linguaggio naturale sui post pubblici del social network *Twitter*. Successivamente abbiamo verificato se e in che modo il *targeting* psicografico fosse applicabile nei social networks, scegliendo come caso studio *Facebook*, con il fine ipotetico di individuare un target potenziale femminile con profilo psicologico coerente con gli ambiti STEM.

I passi che abbiamo compiuto per realizzare quanto sopra esposto sono stati:

1. Individuazione di un dataset idoneo per l'applicazione di tecniche di *Natural Language Processing*[g]con informazioni sulle occupazioni degli utenti;

2. Classificazione del genere e predizione dei tratti di personalità degli utenti del set di dati scelto;

3. Tracciamento delle occupazioni STEM all'interno del dataset e analisi correlazionale degli utenti sulla base del genere, tratti di personalità e della categoria di occupazione;

4. Identificazione e studio di un social network sul quale poter applicare il *targeting* psicografico;

5. Studio di fattibilità del *targeting* psicografico sul social network *Facebook* individuato al punto (4); sotto l'assunzione di possesso di un elenco di utenti di sesso femminile e con i tratti di personalità tipici dell'ambito STEM. Tale assunzione risulta valida in quanto la realizzazione dei punti (2), (3) e (4) ci ha dimostrato che è possibile svolgere *targeting* psicografico (4), con annunci personalizzati, sulla base di genere, tratti di personalità (2) e occupazione degli utenti (3) individuati con tecniche di *Machine Learning*[g]e di analisi del linguaggio naturale.

1.2 Organizzazione del testo

Questa tesi è organizzata come segue. Il Capitolo 2 presenta il contesto oggetto di studio (con l'elaborazione del linguaggio naturale e la teoria della personalità utilizzate per individuare la relazione tra genere, personalità e scelta occupazionale, e la psicografia per svolgere lo studio del *targeting* psicografico). Il Capitolo 3 illustra i lavori già esistenti in letteratura, e che abbiamo utilizzato per ottenere e verificare i nostri risultati. Il Capitolo 4 contiene i materiali, i metodi e i risultati che abbiamo ottenuto, e che ci hanno permesso di rispondere alla domanda *«Il genere di appartenenza e la personalità possono influire sulla scelta occupazionale?»*. Il Capitolo 5 analizza la

pubblicità mirata online, e come questa è integrata all'interno di alcuni social networks (come *Facebook*, *YouTube* e *Instagram*) e motori di ricerca (*Google*); inoltre viene spiegato come sarebbe possibile svolgere il *targeting* psicografico oggetto di questo lavoro su *Facebook*. Al termine, il Capitolo 6 tratta le nostre conclusioni a questo lavoro di tesi e le sue possibili future applicazioni.

Capitolo 2

Contesto

In questo capitolo presentiamo il contesto oggetto del nostro studio che ha coinvolto:

* Elaborazione del linguaggio naturale: utilizzato per comprendere la classificazione del genere e della personalità degli individui da del testo scritto;
* Teoria della personalità: utilizzata per individuare le correlazioni tra i tratti di personalità, genere e scelta occupazionale;
* Psicografia: al fine di comprendere il significato e l'applicazione del *targeting* psicografico.

2.1 Elaborazione del linguaggio naturale

L'analisi del linguaggio con l'impiego della tecnologia, permette di comprendere in modo efficiente e su larga scala come le persone utilizzano e combinano le parole in modo da rilevarne i pensieri, i comportamenti e le emozioni. Inoltre consente, sulla base del lessico d'uso, di prevedere attributi demografici come genere ed età degli individui. Un modo per svolgere tale analisi è mediante testo scritto.

2.1.1 Approccio a vocabolario

L'analisi di un testo scritto può essere effettuata con un approccio a vocabolario:

* **Chiuso:** dizionari o elenchi di parole associati, dai ricercatori, a categorie psico-socialmente rilevanti (come processi sociali e biologici, lavoro ed emozioni). Introdotti per la prima volta nel 1960 e incorporati in programmi informatici che consentono di scansionare automaticamente un testo, di contare la frequenza delle parole di ciascun dizionario e di produrre le frequenze relative; misure utilizzabili anche come variabili in successive analisi statistiche. Alcuni esempi di dizionari, incorporati all'interno di programmi con approccio a vocabolario chiuso, sono *General Inquirer*, *DICTION* e *Linguistic Inquiry and Word Count* (si veda sezione §2.1.1.1).
* **Aperto:** metodi che si basano sull'Informatica, sviluppati dal mondo accademico soprattutto, ma non solo, negli ultimi due decenni (come *Word Embeddings*, maggiori dettagli in sezione §2.1.1.2). Questi sono caratterizzati da algoritmi, che identificano da un insieme di dati linguistici, gruppi di parole semanticamente

correlate in grado di far ottenere informazioni sui campioni e formulare nuove ipotesi sui modelli. Alcuni esempi di approcci a vocabolario aperto sono *Latent Semantic Analysis* e *Latent Dirichlet Allocation* (si veda sezione §2.1.1.2).

Studi recenti [68] definiscono gli approcci a vocabolario chiuso più adatti per verificare in che misura sono presenti precise categorie all'interno di un testo, fornendo così una rappresentazione compatta, oggettiva e stabile dei testi, evidenziando i concetti chiave discussi. Gli approcci a vocabolario aperto, invece, sono un modo efficace per estrarre i temi presenti in uno scritto e il significato delle parole, per chiarire cosa viene discusso nel testo e in che contesto le parole vengano usate. Su questa linea di pensiero si stanno sviluppando nuovi approcci che combinano entrambe le strategie, con l'obiettivo di analizzare i processi psicologici e come si verificano nella vita quotidiana.

2.1.1.1 Approcci a vocabolario chiuso

La metrica principale utilizzata negli approcci a vocabolario chiuso è la frequenza di occorrenza delle parole; soggetta, trattandosi di linguaggio naturale, alla *Zipf's Law* [120] la quale definisce la frequenza di una parola come l'inverso del suo rango (ovvero la posizione occupata dalla parola nel documento/discorso per avere il maggior numero di occorrenze).
Tale metrica permette agli approcci a vocabolario chiuso di identificare due gruppi di parole; le *functions words* (parole di stile come articoli, pronomi e preposizioni) ovvero quelle parole con frequenza maggiore in quanto più comuni in un linguaggio e che occorrono maggiormente, usate in un discorso per definire la struttura sociale e le relazioni [56], e le *content words* (parole di contesto, nomi, verbi regolari, aggettivi e avverbi) che sono invece quelle parole con frequenza inferiore e che per questo occorrono solo in specifiche parti del discorso, descrivendo le emozioni e le caratteristiche di un individuo [56].
Su entrambi i set di parole sono stati sviluppati una serie di dizionari specifici:

* ***General Inquirer*:** sviluppato il 1960 dalla Harvard University [67]. L'ultima versione include 182 dizionari con 8'281 parole uniche, divise in 3 principali sets: 63 *Lasswell dictionaries* (potere e società), 107 *Harvard Psychosociological dictionaries* (psicologia e teorie sociali) e 12 *Stanford Political dictionaries* (interazioni politiche).

* ***DICTION*:** sviluppato il 1980 con lo scopo di analizzare il tono verbale in 500 discorsi presidenziali [114]. *DICTION* assume che nei testi politici ci siano 5 variabili *master* (attività, certezza, comunanza, ottimismo e realismo) e si compone di 31 dizionari *non-overlapping* con 4 variabili, che decodificano la lunghezza delle parole (complessità), il rapporto tra aggettivi e verbi (abbellimento), la frequenza relativa delle parole (insistenza) e il rapporto tra parole uniche e totali (varietà). *DICTION* è stato sviluppato soprattutto per l'impiego nelle politiche commerciali.

* ***Linguistic Inquiry and Word Count***: sviluppato nel 1990 con il fine di analizzare saggi scritti durante gli interventi di scrittura espressiva [39, 148, 149, 140]. *LIWC* è organizzato gerarchicamente, con dizionari che si suddividono in altri sottostanti (esempio, *affective processes dictionary* si suddivide in *positive emotion* e *negative emotion dictionaries*, quest'ultimo che comprende *sadness, anxiety e anger dictionaries*); di conseguenza quando un risultato è correlato a un dizionario di livello inferiore, anche i dizionari ai livelli più alti sono correlati

con il medesimo risultato. *LIWC* ha evidenziato l'importanza dei pronomi nel rilevare diversi processi psicologici; come l'uso della prima persona singolare 'I' correlato con un basso status sociale.

2.1.1.2 Approcci a vocabolario aperto

Gli approcci a vocabolario aperto sono un'alternativa *data-driven* agli approcci a vocabolario chiuso. Tra questi particolare interesse stanno acquisendo gli approcci di *clustering*, capaci di ridurre migliaia di parole in un insieme di variabili ridotto e piú gestibile. Inoltre in letteratura psicologica hanno ricevuto molta attenzione *Latent Semantic Analysis (LSA)* e *Latent Dirichlet Allocation (LDA)*:

* ***Latent Semantic Analysis*:** sviluppata nel 1980 con lo scopo di determinare la somiglianza tra due corpi di testo [22, 23]. *LSA* è un approccio molto simile all'analisi fattoriale (elementi che si allineano in un'unica dimensione all'interno di uno spazio multidimensionale, con conseguente minor numero di fattori latenti) e permette a dei documenti di essere presentati come una combinazione di punteggi fattoriali, i quali clusterizzano le parole che vi appartengono. Di conseguenza le parole che si trovano vicine hanno uno spazio che tende a coincidere e a co-occorrere con le stesse parole nei documenti, questo significa che tali parole sono tendenzialmente correlate. Tale rappresentazione dimensionale permette inoltre di quantificare la distanza semantica tra due parole.
  Tuttavia anche se *LSA* è un metodo robusto in grado di quantificare le differenze semantiche tra i documenti, l'interpretazione delle sue dimensioni è limitata (può capitare che parole caricate sullo stesso fattore siano semanticamente non coerenti); questo perchè la tecnica svolge un'approssimazione del linguaggio come spazio geometrico globale; ignorando totalmente la molteplicità di senso delle parole. Per questo motivo utilizzare *LSA* per l'individuazione su testo scritto di processi psicologici e sociali è limitante, e non sembra una buona idea [68].

* ***Latent Dirichlet Allocation***: *LDA* è un approccio *Generative Probabilistic Clustering* che raggruppa le parole in *topics* o argomenti coerenti di un *corpus* di testo [19]. Gli argomenti possono essere visti come micro-dizionari se paragonati con l'approccio a vocabolario chiuso, tuttavia in questo caso i micro-dizionari vengono generati automaticamente dai dati. Come *LSA* anche *LDA* è un'analisi fattoriale tecnica, però qui l'algoritmo presuppone che ogni parola che occorre possa essere attribuita a uno o più argomenti generati dal *corpus*.
  Il numero di *topics* viene assegnato a priori; le parole vengono assegnate a un *topic* in base alla co-occorrenza con altre parole presenti nel *corpus*; processo reiterato fino a quando non si raggiunge un equilibrio ottimale, ovvero quando tutte le parole del documento sono assegnate a un insieme di *topics* con altre parole semanticamente simili. Tutto ciò viene tradotto in una distribuzione di probabilità a posteriori, che approssima la probabilità di ogni parola che occorre in ogni *topic*. Questi *topics* rappresentano dunque *clusters* di parole, in cui alle parole vengono assegnati dei pesi in base al contributo nel *topic*.
  LDA non presenta gli stessi problemi di *LSA*, in quanto usa una rappresentazione piú strutturata che permette di separare le parole in base al contesto.
  Importante è distinguere la fase di *topic modeling* (costruzione delle variabili) da quella di *topic extraction* (studio delle caratteristiche individuali), che possono essere fatte anche su due datasets distinti. Durante la prima fase avviene la generazione di *topics* da un *corpus* (auspicabile ottenere *topics* di alta qualità e

con coerenza semantica [68]); nella seconda c'è l'etichettatura, sui *topics* modellati nella prima fase, che puó anche essere svolta su un set di dati di dimensioni inferiori al *corpus*.

Altri possibili approcci a vocabolario aperto sono la *Word Embeddings* [104] composta da *embeddings*, parole convertite in vettori, che permettono di catturare le relazioni analoghe tra le parole (esempio in vettori *king - men + woman = queen*) e *Contextual Word Embeddings* [2, 89, 159] in grado di assegnare *embeddings* a dimensioni variabili sulla base del contesto delle parole in analisi.

2.1.2 Creazione di un lessico pesato

Un'altra tecnica che permette l'analisi di un testo scritto è la creazione di un lessico pesato o *weighted lexicon* [129] [131], nata dall'esigenza di rendere pubblico un modello sottoforma di lessico. Difatti prima del lavoro di *Maarten Sap et al.* 2014 [129] numerose erano state le ricerche per prevedere età e genere delle persone da del testo scritto [33]; ma esclusivamente *data driven*, e nessuna finalizzata al riuso e all'approfondimento dei modelli generati. Il concetto del lessico pesato è estendibile anche a campi esterni alla demografia, come predizione di personalità, orientamento politico e analisi del sentimento.
L'approccio di creazione di un lessico pesato coinvolge la definizione di pesi specifici, assegnati a ogni singola parola. L'attribuzione del peso può venire svolta, per esempio, con l'uso di una *support vector machine* (per il genere e l'analisi del sentimento) o una *ridge regression* (per l'età). Inoltre abitualmente il lessico testuale che viene analizzato, anche negli approcci *data driven*, appartiene all'ambito dei social networks.
Per la creazione di un lessico pesato può venire impiegato un approccio univariato, ove si tiene conto di ogni singola parola indipendentemente dalle collinearità, da usare in combinazione a modelli supervisionati; o multivariato quando è il lessico pesato a comporre il modello predittivo stesso e vi è per questo la necessità di tenere conto della covarianza delle parole (ad esempio, chi parla spesso di "capelli" ha più probabilità di parlare di "stile" e "taglio") per evitare conteggi multipli.

2.2 Teoria della personalità

Il termine personalità è legato al termine latino personare, utilizzato per riferirsi agli attori di teatro classico che "parlavano attraverso" la maschera che portavano in scena. Per cui, anche solo nel suo significato, la personalità indica i ruoli che possono venire giocati dalle persone, con modi di fare e sentire, eventualmente coinvolgendo anche gruppi di individui.
In letteratura scientifica la teoria della personalità è definita come un'organizzazione di modi di essere, di conoscere e di agire che assicurano unità, continuità e progettualità alle relazioni dell'individuo con il mondo; implicandone una costruzione che avviene di pari passo con la crescita della persona attraverso le continue interazioni con l'ambiente [24].

Lo studio della personalità avviene sulla base della struttura della personalità, che fa riferimento a come una persona si presenta e si manifesta, e questo avviene con modi di agire (atteggiamenti, comportamenti e sentimenti) uniformi e indipendenti da situazioni e contesto. Un modello adatto allo studio della struttura della personalità

è il modello *Big Five*, il quale fa uso di tratti di personalità per descrivere i modi di agire e le manifestazioni psicologiche degli individui.

2.2.1 Il modello *Big Five*

Il modello a *five-factor (FFM)* conosciuto anche come *Big Five*, è stato sviluppato da un'idea di *Tupes E.C. e Christal R.E. 1961* [144], fondata sugli studi precedenti delle *Cattell's variables* 1940 [15, 43], e a sua volta estesa attorno agli anni 1980 e 1990 da diversi autori [38, 51, 139]. A oggi i *Big Five* rappresentano la tassonomia più completa e ampiamente utilizzata nello studio della personalità grazie alla sua affidabilità, validità e generalizzabilità. I tratti di personalità incorporati sono i seguenti:

* **Amichevolezza:** indica la tendenza di un soggetto verso l'empatia, l'altruismo, la fiducia e la cooperatività [51, 70];
* **Coscienziosità:** indica la tendenza di un soggetto verso il senso di responsabilità, la scrupolosità, la perseveranza, l'autodisciplina e l'affidabilità [51, 70];
* **Estroversione:** indica la tendenza di un soggetto verso la socievolezza, il dinamismo, il bisogno di interazioni sociali e di stimoli esterni [51, 70]. Il lavoro di *Costa P. T. e McCrae R. R. 1980* [139] ha allineato l'elevata estroversione con l'emotività positiva, quest'ultima correlata a sentimenti di felicità, sicurezza personale e soddisfazione.
* **Nevroticismo:** indica la tendenza di un soggetto verso la vulnerabilità, l'instabilità emotiva, l'insicurezza, l'ansia e la depressione [51, 70];
* **Apertura alle esperienze:** indica la tendenza di un soggetto verso l'immaginazione affettiva, la sensibilità estetica, l'attenzione ai sentimenti, la preferenza per la varietà, la curiosità intellettuale e la sfida dell'autorità [51, 70].

Ogni tratto di personalità *Big Five* è una dimensione continua tra due polarità (alto livello contro basso livello). In Tabella 2.1 sono presentati gli aggettivi che descrivono ciascuna di queste polarità [70].

tratto	**alto**	**basso**
Amichevolezza	sensibile, elogiativo, altruista, gentile, bonario, amichevole, cooperativo, piacevole, indulgente, disponibile, fiducioso, caloroso, generoso, dal cuore tenero, affettuoso, apprezzabile, cortese, comprensivo	critico, freddo, ostile, litigioso, duro di cuore, scortese, crudele, irriconoscente
Estroversione	loquace, assertivo, attivo, energico, estroverso, schietto, dominante, entusiasta, esibizionista, socievole, vivace, avventuroso, rumoroso, autoritario	tranquillo, riservato, timido, silenzioso, chiuso, ritirato
Coscienziosità	organizzato, scrupoloso, pianificatore, efficiente, responsabile, affidabile, coscienzioso, preciso, pratico, ponderato, attento	sbadato, disordinato, frivolo, irresponsabile, approssimativo, inaffidabile, smemorato
Nevroticismo	teso, ansioso, nervoso, lunatico, preoccupato, permaloso, pauroso, autocommiserativo, irascibile, instabile, autopunitivo, avvilito, emotivo	stabile, calmo, soddisfatto, non emotivo
Apertura	di ampi interessi, fantasioso, intelligente, originale, perspicace, curioso, sofisticato, artistico, inventivo, acuto, ingegnoso, saggio	banale, di interessi ristretti, semplice, superficiale, poco intelligente

Tabella 2.1: *Aggettivi che definiscono le polarità dei tratti Big Five [70].*

2.2.2 L'Intelligenza Artificiale per prevedere i tratti di personalità

Comunemente per predire i tratti di personalità di una persona, dagli addetti ai lavori, vengono utilizzati dei questionari specializzati [74, 69]; tuttavia già da qualche anno anche l'*Intelligenza Artificiale*[g][95] sta emergendo come strumento adatto a determinare i profili di personalità. Quest'ultima, difatti, si presta bene a decodificarne la *digital footprint* che permette di associare l'individuo a tratti di personalità specifici. Tuttavia può capitare che le persone coltivino, soprattutto nei social networks, profili spesso in contrasto con il loro sé reale; ecco che fare affidamento solo ai dati dei social non è la scelta più adeguata per accertare i tratti della personalità di un soggetto. In quest'ottica una possibile soluzione è l'uso dei dati contestuali; con la consapevolezza, che se il fine ultimo è il *targeting* psicografico, c'è la possibilità che altri fattori, non determinabili a priori, influiscano nella relazione tra tratti della personalità e capacità di persuasione pubblicitaria.

Come si può prevedere la personalità con l'Intelligenza Artificiale?

* Con l'elaborazione del linguaggio naturale, che insieme alla crescita delle comunicazioni digitali sono in grado di fornire un'abbondanza di dati comportamentali delle persone. In tale fronte si possono dedurre le caratteristiche della personalità da informazioni basate sul testo (estratto, ad esempio, dai post nei social networks) usando un approccio a vocabolario aperto [131, 76, 117]. Sembra difatti che i modelli di utilizzo della lingua riflettano la personalità, lo stile di pensiero, le connessioni sociali e gli stati emotivi. In questo modo si è perfettamente in grado di cogliere il comportamento delle persone influenzato da fattori situazionali che verrebbe a mancare con l'utilizzo dei soli social networks [91].
  È possibile anche con l'impiego di *software* già sviluppati, e che utilizzano l'elaborazione del linguaggio naturale, comprendere i tratti di personalità (*IBM* [61]).

* Mediante l'elaborazione dei dati usando un *eye tracker*, il quale rileva i movimenti oculari durante le attività quotidiane che possono predire la personalità [126];

* Misurando l'attività cerebrale, la frequenza cardiaca, la sudorazione e la dilatazione della pupilla come risposta a immagini e video [125].

2.2.2.1 Le *digital footprints*

Le *digital footprints* consistono in tracce di comportamento registrate digitalmente (come, ad esempio, ascoltare musica, cibo o partecipare ad attività finalizzate alla ricerca del brivido) in grado di predire i tratti di personalità della persona a cui appartengono. A oggi gli studi svolti sull'affidabilità delle *digital footprints* non sono uniformi perchè influenzati dalla dimensione del campione, dalla tipologia di *footprints* o dalla piattaforma media utilizzata; per questo l'approccio *gold standard*, per predire la personalità, rimane l'impiego di questionari specializzati. Tuttavia l'uso delle *digital footprints* è una valida alternativa in quei contesti, con limitazioni economiche e di tempo o che richiedono un vasto campione, in cui è impossibile la somministrazione di questionari.

Spinti soprattutto da scopi sanitari e commerciali si stanno facendo strada lo sviluppo di procedure automatizzate; che con l'utilizzo delle *digital footprints* e di un

modello pre-addestrato, permettono lo studio degli effetti delle caratteristiche psicologiche su larga scala [14].
Una volta formato un modello valido è necessario solo l'accesso alle *digital footprints* degli individui, senza l'obbligo di conoscerne le loro caratteristiche psicologiche già collegate durante il *training* del modello, per prevederne i tratti di personalità; consentendo così, vista la grande diffusione dei *Media digitali*[g], la predizione su milioni di persone. Dopodichè su tali predizioni, ad esempio, è possibile astrarre dal livello dell'individuo al livello delle aree geografiche [122] [13] ricercando correlazioni tra i tratti di personalità e le esigenze di una certa fascia di abitanti; oppure sviluppare teorie esplorative guidate dai tratti predetti per definire costrutti di linguaggio della personalità [78] e agevolare così la presa sulla comunità di specifici messaggi sociali.

2.3 Psicografia

Il termine psicografia deriva dalla psicologia e descrive il tentativo di esplorare come le azioni delle persone vengono influenzate dalle diverse personalità umane [134, 138]. Questo, applicato al marketing genera il *targeting* psicografico e consiste nell'assunzione che valori, stili di vita e caratteristiche personali hanno un'influenza significativa sulle decisioni d'acquisto del cliente. Permettendo così un *targeting* più preciso, che fa sentire i clienti indirizzati personalmente dalla pubblicità (o da articoli o notizie create su misura per quella personalità) e più ricettivi al messaggio pubblicitario [44, 108, 50, 91, 146].

2.3.1 *Targeting* psicografico

Il *targeting* psicografico, nato da un rapporto del 1960 (*"Consumer Values and Demands"* pubblicato dallo Stanford Research Institute) e concretizzato da *Arnold Mitchell 1984* in *The Nine American Lifestyles* [106] permette una comunicazione non esclusivamente basata sui dati socio-demografici ma anche su fattori di personalità.
Difatti in una società sempre più diversificata, come la nostra, dove coesistono diversi modelli di vita, il *targeting* socio-demografico risulta inadatto perchè basato solo su dati statistici e per questo non in grado di identificare alcun modello. Inoltre la selezione socio-demografica non certifica l'individuazione di segmenti di pubblico in cui non stanno agendo degli stereotipi.
Dunque quando c'è la necessità di rivolgersi a dei consumatori, intesi nel senso più generale, l'individuazione del gruppo di target deve essere orientata verso aspetti individuali, come la personalità.

2.3.1.1 Composizione e utilizzi

Il *targeting* psicografico viene definito come una composizione di due fasi:

* **Profilazione psicologica:** l'inferenza delle caratteristiche psicologiche di un individuo, che possono includere sia tratti duraturi che stati transitori;
* **Interventi psicologicamente informati:** il tentativo di influenzare i pensieri, le emozioni o i comportamenti dell'individuo facendo appello alle sue caratteristiche psicologiche precedentemente dedotte.

Alcuni studi [60, 107, 16] hanno provato che la comunicazione persuasiva è più efficace se adattata alle caratteristiche psicologiche degli individui; la profilazione psicologica

basata su computer consente di sfruttare questi risultati sul campo e su larga scala. Ad esempio si è riusciti a dimostrare che le vendite possono aumentare fino al +20% se l'aspetto grafico di un sito Web è conforme con l'orientamento motivazionale dominante [59]; anche rivolgersi a un pubblico di personalità diverse sulla base dei loro "Mi piace" su *Facebook* e adattare di conseguenza il contenuto degli annunci, sembrerebbe in grado di incrementare significativamente la probabilità che i potenziali clienti facciano click sull'annuncio [93].
Inoltre si è studiato che accoppiare stati psicologici ed emozioni delle persone in tempo reale possono incrementare l'efficacia degli interventi psicologicamente informati. Questo può venire fatto mediante Sistemi di Raccomandazione consapevoli del contesto, capaci di suggerire i contenuti più rilevanti a un utente in un determinato contesto [83, 115, 165]; oppure in ambito sanitario con interventi adattivi *just-in-time (JITAI)* [153] che sono in grado sfruttando la tecnologia mobile, come telefoni cellulari e sensori, di catturare gli stati emotivi di un paziente e sulla base anche degli stati psicologici precedentemente appresi, offrire interventi personalizzati per migliorarne la salute [111, 90].

2.3.1.2 L'era dei *Big Data*

I *Big Data* hanno rivoluzionato la ricerca nelle scienze sociali, rendendo possibile il passaggio da piccoli studi di laboratorio controllati a studi sul campo su larga scala. Questo è avvenuto anche per il *targeting* psicografico, il quale con l'incremento sia del bacino di dati su cui svolgere la profilazione psicologica che sul quale applicare gli interventi, è risultato applicabile su larga scala nel mondo reale.

I *Big Data* sono caratterizzati dalle 4V (Varietà, Volume, Velocità e Veridicità [13]) e/o dalle 5V (Varietà, Volume, Velocità, Veridicità e Valore [162, 65]):

* **Varietà:** sono numerose le tipologie di dati che possono venire raccolte (da macchine, umani o madre natura), elaborati e gestiti da un sistema. Le scienze sociali tradizionali gestiscono usualmente pochi tipi di dati di provenienza umana; o addirittura anche solo uno, come nel caso dei sondaggi individuali. Invece il *targeting* psicografico, applicabile anche con tracce digitali, è in grado di gestire una vasta gamma di diversi tipi di dati multi-sorgente (come, ad esempio, dati online, musica in streaming, video guardati, immagini pubblicate, testo inviato, metadati sui prodotti acquistati e informazioni geospaziali). Alcuni di questi tipi di dati sono standardizzati e ben strutturati (dati di sondaggi, sensori e metadati), altri si presentano come altamente non strutturati (testo aperto, video e immagini).

* **Volume:** a oggi c'è abbondanza di dati, la maggior parte raccolti da Internet e dai social networks;

* **Velocità:** i dati possono venire raccolti ad alta frequenza e in tempo reale. Inoltre i *Big Data* spesso implicano anche un flusso infinito di dati, il che consente ai ricercatori una visione d'insieme di ciò che stanno osservando.

* **Veridicità**: spesso i dati sono associati a un certo grado di incertezza legata alla loro qualità. I problemi legati alla qualità sono comuni perchè legati alla raccolta in ambiente esterno e non in un laboratorio controllato.

* **Valore:** ci possono essere diversi tipi di vantaggi derivanti dall'elaborazione e dall'analisi dei dati su larga scala. Alcuni di questi possono essere valore monetario, valore sociale e valore di ricerca/istruzione.

Gli scienziati sociali hanno sempre fatto affidamento per i loro studi su tecniche standard, come la correlazione o la regressione lineare, che però risultano inadeguate nel trattamento di grandi quantità di dati. A tal fine, negli ultimi anni, si è assistito a una rapida adozione di metodi provenienti dai campi dell'ingegneria e dell'informatica, capaci di gestire i *Big Data* e produrre *insights* significativi [91].
Tali tecniche si basano soprattutto sul *Machine Learning* il quale può essere supervisionato come regressione e classificazione (per esempio potrebbe venire richiesto a un sistema di saper etichettare una serie di immagini in relazione allo stato d'animo che l'immagine trasmette agli esseri umani) oppure non supervisionato rappresentato, ad esempio, da *clustering* e riduzione della dimensionalità, capace di identificare gruppi di elementi con caratteristiche simili.
Il *Machine Learning* si presenta idoneo a prevedere anche costrutti psicologici significativi come i tratti di personalità [131, 76, 117, 61, 126, 61], da un gran numero di predittori, magari con migliaia di "Mi piace" o tweet di *Facebook*; fatto che nei modelli tradizionali non è possibile, in quanto necessitano di campioni che siano più grandi del numero di predittori. Inoltre i modelli di *Machine Learning* sono capaci di produrre risultati robusti anche quando le variabili non hanno una forma parametrica ben definita, come la distribuzione normale, o quando in presenza di strutture sparse a causa di una bassa densità dei dati in input.

2.3.1.3 Sfide etiche

Il *targeting* psicografico non è uno strumento che fornisce la soluzione perfetta per migliorare il comportamento e la cognizione individuale; ne è da considerarsi un'arma di distruzione della matematica [13]. Difatti la sua utilità e pericolosità per gli individui e la società dipendono dal contesto, cioè da dove, chi e come viene impiegato.

Quando si utilizza il *targeting* psicografico non si può escludere che gruppi di individui con il medesimo tratto di personalità reagiscano in modo diverso all'intervento psicologicamente informato (a causa di gusti personali, interessi o a particolari eventi vissuti); ad esempio non a tutte le ragazze introverse può piacere la programmazione [28], e questo indipendentemente dalle proprie paure e stereotipi della società. Un aiuto nella definizione della soluzione perfetta per migliorare comportamento e cognizione individuale è la combinazione del *targeting* psicografico con il *targeting* socio-demografico; ecco che risulta possibile per esempio prima di cercare l'introversione la selezione di ragazze a cui piacciono i linguaggi di programmazione.
A ogni modo anche tale combinazione presenta delle limitazioni, al raggiungimento della soluzione perfetta, connesse al trattamento della singolarità di un individuo con uno strumento di comunicazione di massa; tali limitazioni possono essere causate, per esempio, dal momento che ogni soggetto del pubblico potenziale sta vivendo (una persona deve essere pronta al cambiamento) o dalle disponibilità monetarie di ognuno, che se presenti possono agevolare l'attualizzazione di una visione d'insieme più ampia.

Il *targeting* psicografico offre il potenziale per rendere gli interventi su larga scala più fattibili ed efficaci; certamente vantaggioso per le aziende in quanto rendere il marketing più efficiente equivale a renderlo più redditizio, con il vantaggio di aiutare anche i consumatori (come, ad esempio, nel caso di sovraccarico delle scelte [25]).

Tale strumento è in grado di creare interventi anche maggiormente accessibili; come osservato con l'uso del *just-in-time* per il trattamento della schizofrenia [18] o nel rilevamento precoce di cambiamenti del comportamento umano [110, 130, 151, 90].
Tuttavia a oggi il *targeting* psicografico non possiede ancora una regolamentazione adeguata e coerente che ne garantisca l'utilizzo nel migliore interesse della società. Ciò comporta che molte delle sfide associate alla profilazione psicologica e agli interventi psicologicamente informati siano difficili da contenere nell'attuale contesto normativo, e potrebbero essere causa di danni sostanziali all'individuo e alla società [13]. Il peggior risvolto che si può ottenere difatti con l'uso improprio di tale strumento è la manipolazione [13, 146], ovvero una forma di persuasione inconsapevole da parte di chi la subisce, basata sulle sue caratteristiche interne e soggettive, che fa si che nella maggior parte dei casi non si stia agendo nel migliore interesse del soggetto [13, 49]. Di conseguenza è la natura stessa del *targeting* psicografico che rende complesso comprendere cosa è manipolativo.
Oltre alla manipolazione c'è anche un'altra eventualità negativa che può scaturire dal *targeting* psicografico; difatti potersi rivolgere a un determinato pubblico sulla base di tratti specifici potrebbe dare luogo a discriminazioni e trattamenti iniqui a livello più generale. Tale situazione di iniquità algoritmica [1] trova la sua maggiore rappresentanza nelle assunzioni dove con il fine di utilizzare la profilazione psicologica per comprendere meglio l'adattamento di un potenziale candidato al lavoro (tuttavia la profilazione psicologica non è l'unico mezzo applicabile per poter parlare di iniquità algoritmica nelle assunzioni, si basti pensare agli effetti della ricerca di competenze digitali su lavoratori over 50 o alla necessità fiscale per le aziende di attivare esclusivamente specifici contratti di apprendistato per giovani lavoratori), si può giungere a risultati con attivi *bias* socio-demografici specifici (come, ad esempio, età o sesso incorrelati dalla personalità).
Un'altra tematica molto importante quanto si tratta di *targeting* psicografico è la privacy. Tradizionalmente la privacy viene definita come l'assenza di osservazione e di disagio di una persona da parte di terzi [13]; tuttavia esistono molteplici definizioni e quella che maggiormente si presta al *targeting* psicografico è l'integrità contestuale. L'integrità contestuale [113] si riferisce all'adeguatezza di un flusso di informazioni personali rispetto alle norme sociali specifiche del contesto. Questo significa che quando il flusso di dati personali è appropriato nel suo contesto, la privacy è salvaguardata, mentre viene violata quando l'uso dei dati personali avviene in contesti inappropriati. Per cui se si considera come buona definizione di privacy l'integrità contestuale, l'attenzione non è più verso quali dati vengono raccolti, ma su come tali dati vengono utilizzati.
La teoria dell'integrità contestuale afferma che devono venire valutati cinque aspetti per determinare se un flusso di informazioni è appropriato: mittente, destinatario, soggetto delle informazioni, tipo di informazioni sull'oggetto e principi di trasmissione in essere (per esempio se c'è stato consenso prima della trasmissione dei dati) [13].
Analizzando il caso di *Cambridge Analytica*[g][146], questa non ha rispettato per nulla il principio di integrità contestuale, violando sia i principi di soggetto che di trasmissione. In quanto, per il sostegno delle proprie campagne di consulenza, la società ha fatto uso di dati personali venduti da terzi (caso *Brexit*, UKIP [146]) e ha prelevato da *Facebook* informazioni personali andando oltre al consenso dato dagli utenti (caso *GSRApp*, sono state raccolte anche le informazioni personali, sugli amici degli utenti dell'applicativo, senza consenso [146]).
Una tutela effettiva e attiva che garantisce la protezione dei dati personali delle persone, da manipolazione e libera circolazione, è il *General Data Protection Regulation* o

GDPR [30]; emanato dal Parlamento Europeo e dal Consiglio dell'Unione Europea in data 27 aprile 2016. Questo stabilisce, nei suoi 173 articoli, che non è possibile alcun trattamento dei dati senza il consenso dell'interessato, che ha con questo dunque pieno potere sulle proprie informazioni, qualunque esse siano; a tutela propria e dell'intera collettività dai rischi insiti nel trattamento dei dati. A garanzia del rispetto del *GDPR* chi si occupa di accedere a dati di terzi (Aziende, Pubblica Amministrazione e Professionisti) deve predisporre e aggiornare un'apposita documentazione che attesti i trattamenti svolti, il rispetto dei diritti da parte degli interessati, la ripartizione dei ruoli e le responsabilità nel trattamento, e le misure di sicurezza implementate informatiche e organizzative. Una qualunque violazione del rispetto della privacy ha come conseguenza da una sanzione pecuniaria (4% del fatturato dell'azienda) fino alla detenzione in carcere, in base alla gravità dell'illecito e alle sanzioni previste dallo Stato membro UE in materia di Garante della Privacy.

Il lavoro di *Ruth E. Appel e Sandra C. Matz 2021* [13] ha cercato di evidenziare i problemi legati all'uso del *targeting* psicografico con le domande seguenti:

1. Quali informazioni sono da considerarsi personali e quali contesti sono da considerarsi privati?
2. In quali contesti l'applicazione del *targeting* psicografico dovrebbe essere limitato?
3. Quali compromessi si possono accettare quando si tratta di processi decisionali algoritmici in cui l'efficienza può essere ottenuta a discapito della giustizia?
4. Quali modelli di business sono da considerarsi accettabili?
5. È corretto che alcune aree come la privacy siano dettate dal mercato?

Le possibili soluzioni individuate [13] che potrebbero aiutare nell'utilizzo proprio ed equo di questo strumento sono:

* **Utilizzare un approccio collaborativo:** lavoro condiviso e in collaborazione tra pubblico e privato, psicologi e informatici, ricercatori e aziende/social media;
* **Migliorare la regolamentazione:** in modo da evitare il ripetersi di casi come *Cambridge Analytica* e per questo integrare la protezione della privacy in modo proattivo, già nella fasi di progettazione, sviluppo, applicazione di tecnologie e sistemi di dati (*privacy by design*). Questo porterebbe la privacy di un individuo a livelli ragionevoli, riducendo il divario tra intenzioni e comportamenti sulla privacy. Inoltre i regolamenti dovrebbero essere più chiari, coerenti e completi; dovrebbero mirare, in particolare, a proteggere i dati in una molteplicità di contesti, tenendo conto di tutti i livelli dei dati e delle informazioni che potrebbero essere generate. È necessario anche che vengano definiti quali dati possono essere utilizzati nel contesto del *targeting* psicografico, determinare quali caratteristiche psicologiche proteggere e specificare quali sono i contesti in cui il *targeting* psicografico può venire utilizzato e/o limitato.

Capitolo 3

Letteratura esistente

Numerosi sono stati i lavori precedenti al nostro che hanno coinvolto lo studio occupazionale, l'elaborazione del linguaggio naturale e la teoria della personalità *Big Five*. In questo capitolo presentiamo le ricerche principali che abbiamo utilizzato per ottenere e verificare i nostri risultati.

3.1 Occupazione

In questa sezione presentiamo alcuni lavori precedenti sull'occupazione.
Il lavoro presentato in sezione §3.1.1 ci ha fornito un dataset idoneo su cui svolgere l'analisi del linguaggio naturale descritta in sezione §4.1.
Il lavoro presentato in sezione §3.1.2 ci ha fornito le informazioni necessarie per individuare le occupazioni STEM descritte in sezione §4.5.
Il lavoro presentato in sezione §3.1.3 è stato un utile riferimento durante l'analisi dei risultati di predizione della personalità descritta in sezione §4.3.3.1, in quanto ci ha permesso di approfondire la coscienziosità, rilevata nel nostro dataset di lavoro, in rapporto con la sovraqualificazione dei lavoratori.

3.1.1 Analisi delle occupazioni di utenti *Twitter*

L'obiettivo del lavoro [119] è stato quello di predire le classi occupazionali di un sottoinsieme scelto di profili di utenti pubblici di *Twitter* composto da persone fisiche e aziende. Per mappare le occupazioni degli utenti gli autori hanno preso a riferimento lo *Standard Occupational Classification (SOC)* [112], sistema di classificazione occupazionale istanziato dal governo del Regno Unito e sviluppato dall'*Office for National Statistics (ONS)*[g].

3.1.1.1 Creazione del dataset

Per creare *Twitter Occupation Dataset* i ricercatori *Preoţiuc-Pietro et al. 2015* [119] hanno nell'ordine compiuto i seguenti passi:

1. Valutato, con annotazione del campo descrizione dell'utente, la percentuale di profili *Twitter*, persone fisiche e aziende, con una chiara menzione alla loro occupazione;

2. Scelto tra quanto risultato in (1) un campione di un 1% di utenti la cui cronologia conteneva almeno 200 tweets e la maggioranza in inglese;

3. Ricercato nel campo descrizione degli utenti di (2) i titoli di lavoro dichiarati conformi alla classificazione *SOC*, con a seguire svolta l'aggregazione degli utenti in categorie da *3-digit*;

4. Proceduto con la rimozione di ambiguità (per esempio, nessuna descrizione, utenti senza corrispondenza con la categoria di occupazione assegnateli, o assegnazioni a categorie multiple). Questo passo ha comportato la creazione di un dataset con 10'796'836 tweets e 5'191 utenti sia persone fisiche che aziende.

5. Utilizzato un set di dati separato come *corpus* di riferimento. A tal fine è stato scelto dai ricercatori un estratto di *Twitter Gardenhose stream*, campione rappresentativo del 10% dello stream di *Twitter*; sulle quali parole hanno applicato tokenizzazione e filtraggio per la lingua inglese, generando così un *corpus* finale composto da 71'555 parole.

3.1.1.2 Predizione delle occupazioni

Per la predizione delle occupazioni gli autori di [119] hanno generato e testato due classi di *features*: a livello d'utente che si sono dimostrate non utili, nel prevedere la classe di lavoro, e testuale che invece si sono presentate adatte allo scopo. Nel primo caso si trattavano di informazioni aggregate dell'utente o statistiche sui tweets; invece le *features* testuali derivavano dall'insieme aggregato di tweets dell'utente, e rappresentavano ciascun utente con la loro distribuzione calcolata sul *corpus* di riferimento. Per ottenere le *features* testuali gli autori hanno fatto uso di tecniche di *embeddings* e di *cluster words*, che gli hanno permesso di lavorare con matrici di somiglianza e gruppi di parole correlate. Anche in questo caso sono state riscontrate delle differenze di prestazioni che impattavano sulla predizione finale; i *clusters* hanno evidenziato delle prestazioni superiori rispetto agli *embeddings*, probabilmente perchè questi ultimi enfatizzavano troppo le parole comuni.
Per la predizione di classe è stata utilizzata la classificazione con processi gaussiani; che ha permesso di fare inferenza sulla distribuzione delle occupazioni per ciascun utente e di, in combinazione con la *logistic function* [71], individuare la classe di occupazione di appartenenza dell'utente. Tuttavia tale predizione, individuata sottoforma di *likelihood*, era di forma non gaussiana e per questo non trattabile. Per risolvere il problema si è fatto uso dell'*Expectation Propagation* [105]; la quale ha approssimato la formulazione a posteriori della *likelihood* a una gaussiana, rendendo la predizione trattabile e di conseguenza permettendo ai ricercatori di assegnare a ogni rappresentazione dell'utente la classe occupazionale con *maximum likelihood* [21].

3.1.2 Definizione delle occupazioni STEM

L'obiettivo della *UK Commission for Employment and Skills*[g], autore del lavoro [40], è stato quello di definire le occupazioni STEM basandosi sulle *high level STEM skills*; ovvero alti livelli di competenze STEM dipendenti dal tipo di occupazione e dalla domanda di competenze nel mercato del lavoro. Durante l'analisi preliminare delle *high level STEM skills* nel mercato del Regno Unito, è stato notato dalla Commissione come non sempre il possesso di specifiche competenze era nella realtà associato a lavoratori con un'elevata formazione (*Livello 4+*[g]); ad esempio nel caso di tecnici elettronici ed elettrici solo il 10% è risultato in possesso di una qualifica adeguata.

3.1.2.1 Classificazione in STEM/NOSTEM

In [40] per definire le occupazioni STEM gli autori inizialmente hanno creato una *initial longlist* di tutte le occupazioni STEM/NOSTEM. Per tale scopo sono stati impiegati i dati dei laureati e lavoratori della *Labour Force Survey* [62] ed è stato sviluppato un indice combinato per *Numeracy*[g]e *Problem Solving*[g]; dopodichè con l'uso di entrambe le metriche è stato applicato il *k-medians cluster (con k=2)* [73] sulle occupazioni decodificate utilizzando lo standard *SOC*.
Come secondo passo, i membri della Commissione hanno proceduto con la ridefinizione della lista iniziale con:

* Rimozione dei falsi positivi chiari (per esempio, le Forze Armate);
* Esclusione delle professioni medico-sanitarie;
* Eliminazione delle professioni di insegnamento e basate sulla matematica ma non esplicitamente legate a scienze, ingegneria e tecnologia;
* Rimozione delle occupazioni che necessitano di un'alta competenza STEM ma che hanno un focus differente (per esempio, Piloti);
* Non considerazione del lavoro di modellizzazione (per esempio, Tessitori e Professionisti associati alla conservazione dell'ambiente) perchè con dati statistici insufficienti a supportarne l'inclusione. Di contro sono state incluse occupazioni di gruppi di unità tecniche non classificate come STEM per mancanza di dati.

3.1.2.2 Analisi occupazionale

Nel report prodotto la *UK Commission for Employment and Skills* ha impiegato i risultati della classificazione delle occupazioni per tracciare un profilo delle caratteristiche principali delle occupazioni STEM e compiere una valutazione del fabbisogno del mercato del lavoro associato alle professioni STEM di alto livello, con analisi della copertura occupazionale dell'apprendistato.
In conclusione i membri della Commissione hanno stabilito come le *high level STEM skills* sono di fondamentale importanza per l'economia britannica in termine di occupazione, produttività, innovazione e competitività. Tuttavia vi è una carenza di persone qualificate in specifiche aree professionali, come nel caso dei tecnici, e una formazione da parte dei soli datori di lavoro può non risultare sufficiente a colmare la carenze e soddisfare le esigenze delle imprese.

3.1.3 Sovraqualificazione

Nel lavoro [79] *S. LaRochelle-Côté e D. W. Hango 2016* hanno misurato la sovraqualificazione presente nei lavoratori canadesi, tra i 25 e i 64 anni, in possesso di un titolo universitario. I ricercatori si sono occupati anche di stimare quanto i lavoratori qualificati si sentissero insoddisfatti del proprio lavoro.
Dai risultati ne è emersa una percentuale di occupazione maggiore tanto più elevata era la qualifica in possesso dei lavoratori; tuttavia quando un lavoratore ricopriva una posizione inferiore all'istruzione ricevuta, si è osservato che questo impattava negativamente sulla soddisfazione, mantenimento del posto e relativi guadagni dello stesso. Gli autori si sono occupati anche di misurare le sovraqualificazioni presenti nel mercato d'occupazione canadese; a tal fine hanno analizzato le categorie STEM, Umanistiche, delle Scienze sociali, di Insegnamento e della Salute. La sovraqualificazione in ambito

STEM osservata è stata del 30% circa, al terzo posto dopo le facoltà Umanistiche (44.5%) e le Scienze sociali (33.2%).

3.2 Elaborazione del linguaggio naturale

In questa sezione presentiamo alcuni lavori precedenti sull'elaborazione del linguaggio naturale.
I lavori presentati nelle sezioni §3.2.1.1 e §3.2.1.2 ci hanno permesso, con modello prodotto e approcci, di svolgere la predizione e l'analisi dei tratti di personalità descritte in sezione §4.3.
Il lavoro presentato in sezione §3.2.1.3 lo abbiamo impiegato per l'approccio del lessico pesato e il modello prodotto per la classificazione del genere descritto in sezione §4.2.

3.2.1 Approcci sul testo scritto per personalità, genere ed età di utenti *Facebook*

I lavori [129, 131, 76] sono molto simili negli obiettivi. In ciascuno è stata svolta l'analisi del linguaggio naturale su post di *Facebook* di utenti che avevano fatto uso dell'applicazione *myPersonality*[g]; su questi sono stati studiati il genere, l'età [129, 131, 76] e la personalità [131, 76]. Gli approcci testati sono stati a vocabolario chiuso [131] (si veda sezione §2.1.1.1, in [76] solo a fine di verifica dell'approccio a vocabolario aperto), vocabolario aperto [131, 76] (si veda sezione §2.1.1.2) e la creazione di un lessico pesato [129] (si veda sezione §2.1.2).

3.2.1.1 Vocabolario chiuso

Il lavoro di *Schwartz et al. 2013* [131] ha testato sui post un approccio a vocabolario chiuso. Tale tipo di approccio associa il linguaggio usato da un individuo a una categoria, per questo è stato indicato dagli autori anche come *word-count approach*, e lo hanno definito con la formula 3.1:

$$p(category|subject) = \frac{\sum_{word \in category} freq(word, subject)}{\sum_{word \in vocab(subject)} freq(word, subject)}, \tag{3.1}$$

dove $freq(word, subject)$ è il numero di volte in cui un soggetto utilizza la parola $word$ e, al denominatore, $vocab(subject)$ indica l'insieme di tutte le parole utilizzate dal soggetto.

Le categorie utilizzate nell'approccio [131] sono state le categorie *LIWC* [140] (si veda sezione §2.1.1.1) e il *subject* gli utenti di *myPersonality*. Dopodichè per collegare le categorie con i tratti di personalità, genere ed età i ricercatori hanno utilizzato la regressione dei minimi quadrati ordinari. Il coefficiente della variabile esplicativa, in categoria *LIWC*, è stata considerata la forza della relazione.

3.2.1.2 Vocabolario aperto

Nei lavori *Schwartz et al. 2013* [131] e *Margaret L. Kern et al. 2014* [76] viene presentata la tecnica di *Differential Language Analysis (DLA)*. Questa si basa su tre aspetti.

1. **Estrazione delle caratteristiche linguistiche:** in cui le caratteristiche linguistiche vengono esaminate su parole, frasi e argomenti.
   Per parole e frasi sono state esaminate, in entrambi i lavori, sequenze di n-grammi da 1 a 3 parole. Ogni parola è stata individuata con l'uso di un *tokenizer* sviluppato dagli stessi autori con l'uso di *"happyfuntokenizing" di Pott* [118], che ha permesso la cattura anche di emoticon. Inoltre sono state mantenute solamente le parole delle frasi con un alto valore informativo. A tal scopo i ricercatori hanno eseguito il calcolo di *pointwise mutual information*, formula 3.2, cioè il rapporto tra la probabilità condizionale e la probabilità indipendente di osservare la frase:

$$pmi(phrase) = log \frac{p(phrase)}{\prod_{word \in phrase} p(word)}. \tag{3.2}$$

   Sono state rimosse tutte le frasi con *pmi* inferiore al doppio del numero di parole contenute nella frase, in modo da mantenere solo le parti significative del discorso. La normalizzazione dei conteggi delle parole e frasi, formula 3.3, è stata svolta sulla base del numero totale di parole usate da ciascun soggetto ($vocab(subject)$) e applicata l'*Anscombe trasformation* per stabilizzare la varianza, formula 3.4:

$$p(phrase|subject) = \frac{freq(phrase, subject)}{\sum_{phrase` \in vocab(subject)} freq(phrase`, subject)}, \tag{3.3}$$

$$p_{ans}(phrase|subject) = 2\sqrt{p(phrase|subject) + \frac{3}{8}}. \tag{3.4}$$

   Per i *topics* o argomenti l'approccio dei ricercatori [131] è stato quello di creare gruppi di parole utilizzando la tecnica di *Latent Dirichlet Allocation (LDA)* [19] (si veda sezione §2.1.1.2). I post di *Facebook* sono stati interpretati come contenitori di *topics* e i *topics* composti da una distribuzione di parole. Inoltre essendo che le parole erano note, la variabile latente argomenti è stata stimata con *Gibbs Sampling*. A fini pratici quanto appena spiegato si è tradotto per [131] nell'utilizzo dell'implementazione *LDA* di *Mallet package* [94].
   In questo modo è stato possibile calcolare la probabilità che un soggetto utilizzi un *topic* come:

$$p(topic|subject) = \sum_{word \in topic} p(topic|word) * p(word|subject), \tag{3.5}$$

   dove $p(word|subject)$ è la parola normalizzata utilizzata dal soggetto (formula 3.3) e per $p(topic|word)$ si intende il risultato della procedura *LDA*, ed è la probabilità del *topic* data la parola.

2. **Analisi correlazionale:** con regressione dei minimi quadrati ordinari. In [131, 76] la variabile esplicativa *target* è la forza della correlazione e le covariate (variabili sul quale viene misurata con la correlazione la dipendenza verso l'esplicativa) le caratteristiche linguistiche, l'età e il genere.
   Una correlazione è una relazione tra due variabili espressa da un coefficiente di correlazione, ovvero un numero compreso tra -1 e 1; che indica la direzione (il segno) e la forza (il valore assoluto) della relazione. Quando il segno del coefficiente assume valore positivo, ovvero vi è correlazione positiva tra le variabili, significa che all'aumento di una delle due variabili aumenta anche la seconda; se invece il coefficiente è negativo significa che vi è correlazione negativa e all'aumentare della

prima variabile, la seconda cresce ma di segno opposto; infine se il coefficiente è nullo non vi è alcuna correlazione. Il valore assoluto della correlazione, invece, denota l'entità della correlazione tra le variabili; maggiore è il valore assoluto più forte è la correlazione.

3. **Visualizzazione:** mediante *word clouds*. In [131, 76] è stata utilizzata la correlazione come dimensione della parola e la frequenza per il colore. I ricercatori hanno scelto di non costruire esclusivamente le *word clouds* sulla base della frequenza in quanto pratica da studi precedenti [58] criticata sotto gli aspetti di analisi testuale, contesto dei dati e narrazione. Questo perchè l'utilizzo della sola frequenza non permette di individuare alcuna correlazione tra le parole ("child" e "death" possono avere entrambe un'alta frequenza in uno specifico contesto, ma questo non implica un rapporto tra le due), inoltre può dare un peso eccessivo a termini che in base al contesto di analisi non hanno alcun valore ("the" è una parola utilizzata molto nel linguaggio, tuttavia non ha alcun significato se il contesto è quello della personalità perchè d'occorrenza in tutti i tratti); tutto ciò rende impossibile una narrazione veritiera dei risultati di una *word clouds* costruita esclusivamente sulla frequenza delle parole. Il *tool* utilizzato per la loro generazione è stato *Wordle* [147].

3.2.1.3 Creazione di un lessico pesato

Nel lavoro di *Sap et al. 2014* [129] viene presentato il metodo di creazione del *weighted lexicon*; ottenuto usando i coefficienti dei modelli di regressione lineare multivariata e di classificazione.
La formula di *weighted lexicon* 3.6 è stata definita dagli autori come segue e se applicata ai modelli di regressione lineare multivariata, formula 3.7, permette la classificazione del genere e la predizione dell'età:

$$usage_{lex} = \sum_{word \in lex} w_{lex}(word) * \frac{freq(word, doc)}{freq(*, doc)}, \tag{3.6}$$

dove $w_{lex}(word)$ è il peso della parola nel lessico lex, $freq(word, doc)$ è la frequenza della parola sul documento e $freq(*, doc)$ è la somma di tutte le frequenze nel documento.

$$y = (\sum_{f \in features} w_f * x_f) + w_0, \tag{3.7}$$

dove x_f è il valore della *feature* ($\frac{freq(word,doc)}{freq(*,doc)}$ se le *features* hanno tutte parole con frequenze relative), w_f il *feature coefficient* (ovvero $w_{lex}(word)$) e w_0 l'intercetta.

Il metodo di *weighted lexicon* illustrato in [129] consiste nel far apprendere il coefficiente di ciascuna parola (w_{lex}) a una *ridge regression* per l'età o a una *support vector machine* per il genere. Successivamente, con modelli aggregati a livello di utente e la parola aggregata come frequenza relativa, i ricercatori hanno utilizzato il lessico pesato, formula 3.6, per predire genere ed età degli utenti.
La creazione di un lessico pesato è stato valutato anche su social networks ulteriori oltre a *Facebook* (come *Bloggers*, *Twitter* e in combinazione) riscontrando per l'età una MAE (*Mean Absolute Error*, metrica di prestazione con media degli errori di un modello di regressione) abbastanza bassa, tra i valori 3-11 in anni, e per il genere un'accuratezza media tra 80%-90%.

3.3 Teoria della personalità *Big Five*

In questa sezione presentiamo alcuni lavori precedenti sulla teoria della personalità *Big Five*.
I lavori presentati nelle sezioni §3.3.1, §3.3.2, §3.3.3 e §3.3.4 sono stati impiegati come punto chiave nell'analisi correlazionale di personalità delle occupazioni descritta in sezione §4.5.3.
I lavori presentati in sezione §3.3.5 sono stati utilizzati durante l'analisi dei risultati di predizione della personalità descritta in sezione §4.3.3.1.

3.3.1 Nevroticismo e ansia per la matematica negli studenti STEM

L'obiettivo del lavoro [87] è stato quello di analizzare il ruolo del nevroticismo e dell'ansia per la matematica (*math anxiety*) su un campione di studenti universitari STEM, laureandi sia maschi che femmine (70+70) e con pari livello di *Quoziente Intellettivo (QI)*[g].
Lo studio ha raccolto misure di *math anxiety* e nevroticismo mediante la somministrazione di questionari; inoltre sono state indagate le capacità di calcolo (*Numeracy*) degli studenti attraverso un test di valutazione delle competenze.

Il lavoro di *Maristella Lunardon, Tania Cerni e Raffaella I. Rumiati 2022* [87] ha ottenuto i seguenti risultati:

1. I maschi hanno superato le femmine nel test di calcolo;
2. Le femmine hanno ottenuto punteggi più elevati rispetto ai maschi sia per *math anxiety* che per il nevroticismo;
3. La *math anxiety* è stata associata positivamente solo nei maschi, e negativamente associata al calcolo solo nelle femmine;
4. Nelle femmine il punteggio di calcolo è stato positivamente correlato al nevroticismo.

Le studentesse STEM analizzate sono risultate essere più nevrotiche e ansiose in matematica dei colleghi maschi (2), e hanno ottenuto prestazioni inferiori nel calcolo (1) durante i compiti. Il nevroticismo tuttavia non ha mostrato un impatto negativo, ma positivo, in termini di performance nelle femmine (4); cosa che invece non è accaduta per l'ansia (3). Per gli autori di [87] questo è indicatore che se il soggetto è altamente nevrotico, in una situazione di forte ansia la sua performance tende a salire, perchè maggiormente in linea con il proprio tratto psicologico.
Invece nei maschi l'ansia per la matematica non ha mostrato avere alcun impatto significativo sulla performance del calcolo; anzi rispetto alle femmine si sono dimostrati meno nevrotici e capaci di ottenere valutazioni nei test superiori.

Il lavoro [87] non è l'unico che ha indagato i tratti psicologici degli individui STEM, ne sono un esempio le ricerche compiute in [28, 27, 127, 124, 103].
Nel lavoro di *Johan Coenen, Lex Borghans e Ron Diris 2021* [28] sono stati esaminati gli studenti STEM. Le femmine STEM sono state valutate come maggiormente aperte alle esperienze; invece i maschi STEM tendenti all'introversione. In generale gli autori hanno rilevato che uno studente STEM era più aperto, ma meno estroverso e meno

amichevole di uno studente NOSTEM.
Nel lavoro di *Deborah A. Cobb-Clark e Michelle Tan 2011* [27] sono state esaminate le occupazioni STEM. Sono stati riscontrati uomini in correlazione negativa con coscienziosità ed estroversione, ma positiva con amichevolezza, apertura alle esperienze e stabilità emotiva, e donne che hanno presentato le medesime correlazioni dei colleghi tuttavia meno marcate negativamente.
I lavori [87] e [27] hanno ottenuto risultati opposti in termini di nevroticismo STEM, con femmine che anzi sono più stabili emotivamente dei maschi [27]. Tali contrapposizioni sono giustificate dalle differenze tra i campioni esaminati (studenti e occupazioni) e alle tecniche di predizione della personalità *Big Five* utilizzate dai diversi gruppi di ricerca (somministrazioni di questionari con scale differenti).
Nei lavori [127, 124, 103] vengono esaminate alcune specifiche discipline appartenenti all'ambito STEM.
In *Norsaremah Salleh et al. 2010* [127] i ricercatori hanno riscontrato come i tratti di personalità che caratterizzano maggiormente un gruppo di studenti neozelandese, durante i loro studi di programmazione, erano il nevroticismo e l'amichevolezza; mentre le correlazioni di segno opposto sono state ottenute con estroversione e apertura alle esperienza. Gli autori hanno anche notato che il 75% degli studenti oggetto della loro analisi era di sesso maschile; ipotizzando dunque che questo limitasse i loro risultati e la valutazione di performance su coppie di studenti con diverso grado di nevroticismo.
Gidi Rubinstein 2002 [124] ha analizzato i tratti di personalità *Big Five* di studenti iscritti a diverse facoltà: Legge, Scienze sociali, Scienze naturali e Arte. Per quanto riguarda la facoltà di Scienze naturali (comprensiva dei corsi di laurea in Matematica, Informatica, Fisica e Chimica) l'autore ha riscontrato una correlazione positiva generale di tutti e cinque i tratti con donne che, come già accaduto nel lavoro [27], presentavano correlazioni positive più marcate dei colleghi uomini. Tuttavia questa volta la correlazione positiva è stata con il nevroticismo [87] e non più con la stabilità emotiva.
Meier Michaela, Vogel Stephan e Grabner Roland 2021 [103] hanno analizzato i tratti di personalità di studenti con diversi gradi di esperienza in matematica. Ne è risultato che gli studenti di Matematica con elevata esperienza erano meno nevrotici e più amichevoli rispetto a chi non era un matematico.
In conclusione il lavoro [127] ancora una volta sembra distanziarsi dal lavoro [87]; invece il lavoro [124] si mostra in linea con le conclusioni sul nevroticismo femminile. Questi risultati possono forse indicare che in ambito STEM il nevroticismo è femminile; tuttavia in Informatica il nevroticismo è correlato fortemente con il sesso maschile e originato dalla programmazione [124]. Di conseguenza i tratti di personalità ci appaiono dipendenti dalla distribuzione del campione analizzato all'interno dei corsi di laurea STEM.

3.3.2 Correlazioni tra i tratti di personalità *Big Five*

Il lavoro *Azeem Akbar, Amara Malik e Nosheen Fatima Warraich 2023* [10] analizza le correlazioni tra i tratti di personalità *Big Five* di professionisti che lavoravano presso biblioteche universitarie in Pakistan. Per analizzare ciascun tratto è stato somministrato ai partecipanti un questionario online. I risultati che ne sono conseguiti hanno evidenziato un'elevata correlazione positiva tra coscienziosità e amichevolezza ed un'alta correlazione negativa, invece, tra amichevolezza e nevroticismo.

3.3.3 Tratti di personalità per genere

Lavori come [72, 161] hanno indagato, in diverse forme, l'incidenza dei tratti di personalità *Big Five* in base al genere di appartenenza.
A. F. Jorm 1987 [72] ha riscontrato una correlazione positiva tra il genere femminile e il nevroticismo più alta rispetto ai maschi. Inoltre ha legato le differenze di sesso e di età del nevroticismo alla depressione.
Weisberg YJ, Deyoung CG e Hirsh JB 2011 [161] hanno osservato ulteriori correlazioni oltre al nevroticismo per le femmine come l'apertura alle esperienze, l'estroversione, l'amichevolezza e la coscienziosità; indicatori che le donne sono correlate maggiormente con l'alta polarità presente in ciascun tratto di personalità.

3.3.4 Tratti di personalità in un utente dei social networks

In lavori come [145, 158, 75] è stato studiato il legame tra personalità e social networks.
In *Yusuke Umegaki e Ayaka Higuch 2022* [145] è stato osservato che gli utenti di *Twitter* ottenevano, rispetto a chi non ne faceva uso, punteggi inferiori in estroversione, amichevolezza, coscienziosità e apertura, e significativamente superiori nel nevroticismo. In generale su tutte le piattaforme analizzate (*Twitter*, *Instagram* e *Facebook*) i ricercatori hanno riscontrato un nevroticismo maggiore degli utenti rispetto ai non utenti.
Anche in *Amichai-Hamburger et al. 2002* [158] è risultato che chi faceva uso del mondo *real-me*, ovvero impiegava chat e Internet per comunicare con gli amici, era tendenzialmente meno estroverso e più nevrotico rispetto a chi utilizzava l'approccio faccia a faccia. Inoltre i ricercatori hanno osservato che le femmine *real-me* anche se correlate con l'introversione, lo erano di meno dei colleghi maschi; tuttavia erano maggiormente correlate al nevroticismo.
Rachubińska K. et al. 2021 [75] ha analizzato nello specifico cosa accade nelle donne che utilizzano il social network *Facebook*. Qui i ricercatori hanno rilevato che le femmine utilizzatrici erano in quota minoritaria nevrotiche; tuttavia nel momento in cui presentavano alti tratti di nevroticismo avevano un'alta probabilità di sviluppare dipendenza. I tratti di personalità, maggiormente rilevati tra le donne utilizzatrici del social network, sono stati in ordine decrescente coscienziosità, amichevolezza, estroversione, apertura alle esperienze e nevroticismo.

In conclusione i lavori [145, 158] supportano l'ipotesi che chi utilizza un social network abbia dei tratti di personalità più marcati in nevroticismo e introversione, rispetto a chi non ne fa uso. Inoltre per le donne con alto tratto di nevroticismo [75] non viene evidenziata una grande attrazione nell'uso dei social networks.

3.3.5 Tratti di personalità in un buon lavoratore

La coscienziosità è uno dei tratti che se a livelli moderati porta numerosi benefici all'essere umano, sia in termini di salute che di qualità della vita. Inoltre per avere successo in ambiente lavorativo è il tratto che sembra essere indispensabile per un lavoratore. Questi aspetti sono stati analizzati da diversi lavori, come [150, 155, 41].
In *Roberts B. W. 2014* [150] la coscienziosità è risultata essere il tratto di personalità *Big Five* determinante per la salute, l'invecchiamento positivo e il capitale umano. I ricercatori hanno osservato che le persone coscienziose assumevano comportamenti simili, come laboriosità, ordine, responsabilità verso gli altri, autocontrollo e rispetto delle regole.

In *Michael P. Wilmot e Deniz S. Ones 2019* [155] il gruppo di ricerca ha evidenziato una relazione stretta tra coscienziosità e performance nel lavoro, incentivata dalla motivazione e dal proseguimento di un obiettivo.
Piero Esposito e Sergio Scicchitano 2022 [41] hanno in aggiunta individuato una correlazione tra coscienziosità e la soddisfazione in ambito lavorativo, e anche una correlazione positiva con sovraeducazione e sovraqualificazione.

Capitolo 4

Materiali, Metodi e Risultati

In questo capitolo presentiamo i materiali, i metodi e i risultati che abbiamo ottenuto e che ci hanno permesso di rispondere alla domanda *«Il genere di appartenenza e la personalità possono influire sulla scelta occupazionale?»*.

4.1 *Twitter Occupation Dataset*

Il dataset che abbiamo impiegato per l'analisi del linguaggio naturale è stato *Twitter Occupation Dataset* [119] (si veda sezione §3.1.1) in cui per ogni utente abbiamo predetto il genere e i tratti di personalità *Big Five*. Con queste informazioni siamo riusciti a individuare le correlazioni tra genere, personalità e categorie occupazionali. Inoltre abbiamo realizzato delle *word clouds* per studiare le parole più utilizzate da maschi, femmine e in base alla personalità per verificare la ragionevolezza dei modelli e degli approcci utilizzati.

4.1.1 Struttura

Twitter Occupation Dataset è strutturato come di seguente:

* `jobs-tweetids`, con identificativo utente e identificativo del tweet;
* `jobs-unigrams`, per ogni identificativo utente viene costruita la *Bag of words (BOWs)* associata all'utente, composta da id della parola e frequenza;
* `dictionary`, contiene il *corpus* di riferimento composto dall'associazione univoca tra id parola e parola;
* `job-users`, per ogni identificativo utente è espresso il suo codice occupazionale in *3-digit SOC code*;
* `keywords`, per ogni codice occupazionale viene indicata la descrizione della classe e le parole chiavi dei lavori che identificano la categoria di utenti.

4.1.2 *Bag of words*

Bag-Of-Words Dataset (BOWs Dataset) è un dataset che ci siamo costruiti in cui ogni entry è stata realizzata sulle singole *Bag of words* degli utenti di *Twitter Occupation*

Dataset.
BOWs Dataset è stata la nostra base di partenza con cui abbiamo svolto la predizione del genere, dei tratti di personalità e la realizzazione delle *word clouds*.

4.1.2.1 Metodologia

Per ogni *BOWs* degli utenti, abbiamo estratto ciascuna parola e calcolato x_{word}, formula 4.1; che rappresenta la *term frequency*, ovvero la frequenza relativa di una parola rispetto alla *BOWs* dell'utente:

$$x_{word}(user) = \frac{freq(word, BOWs_{user})}{freq(*, BOWs_{user})}. \tag{4.1}$$

In formula 4.1 il nominatore $freq(word, BOWs_{user})$ è la frequenza della parola *word* nel *BOWs* di un utente *user* e il denominatore $freq(*, BOWs_{user})$ rappresenta la somma di tutte le frequenze delle parole contenute nella *BOWs* dell'utente.
Questo ci ha permesso di realizzare *Bag-Of-Words Dataset*, con una entry per ogni utente e le seguenti colonne:

* Identificativo utente;
* *BOWs* dell'utente espressa come *"word" : x_{word} : freq*, per ogni parola ivi contenuta.

4.1.3 Personalità e occupazione

Personality-Occupation Dataset (PO Dataset) è un dataset che ci siamo costruiti in cui le entries sono state realizzate con le informazioni di genere, personalità e di occupazione degli utenti di *Twitter Occupation Dataset*.
PO Dataset è stata la nostra base per l'analisi occupazionale.

4.1.3.1 Metodologia

Per ogni utente abbiamo collegato tra loro le informazioni in nostro possesso da *Twitter Occupation Dataset*, che abbiamo ottenuto dalla classificazione del genere descritta in sezione §4.2 e la predizione della personalità *Big Five* descritta in sezione §4.3.
Questo ci ha permesso di realizzare *Personality-Occupation Dataset*, con una entry assegnata a ciascun utente e le seguenti colonne:

* Genere predetto per l'utente;
* Correlazione normalizzata di ciascun tratto di personalità *Big Five* dell'utente;
* Alti/bassi tratti di personalità predetti, se ve ne sono, per l'utente;
* Descrizione dell'occupazionale dell'utente;
* *SOC code* occupazionale dell'utente (*UK SOC code*).

4.2 Genere

Per la classificazione del genere abbiamo deciso di istanziare il modello *emnlp14gender* di *Age and Gender Lexica* [129] (si veda sezione §3.2.1.3) il quale si compone dei

parametri *term* e *weight*, che associano ciascuna parola del modello a un genere.
In dettaglio, Tabella 4.1, *term* sono parole estratte dai post di *Facebook* degli utenti che hanno utilizzato l'applicazione *myPersonality*; *weight* è il *feature coefficient* in cui w_0 è l'intercetta. Quando il coefficiente ha valore positivo allora si tratta di un termine più comunemente utilizzato da femmine, altrimenti da maschi.

term	**weight**
_intercept	-0.06724152
raining	-29.5501115991
yellow	-14.3996802779
four	9.06714906319
gag	18.8233764377
woods	-37.2521575792
hanging	61.2268981744
increase	-26.9592449062
electricity	158.279682173
funk	-67.9061884798
lord	16.9656445271

Tabella 4.1: *Modello emnlp14gender.*

Il modello è stato impiegato per estrarre i pesi da assegnare alle parole delle *Bag of words* e applicare con questi la formula di *weighted lexicon* (si veda sezione §3.2.1.3). Tale procedura ci ha consentito di classificare ciascuna *BOWs* di *Twitter Occupation Dataset* appartenente a un utente femmina oppure a un maschio.

4.2.1 Metodologia

Per classificare il genere degli utenti contenuti in *BOWs Dataset* abbiamo fatto riferimento all'approccio del lessico pesato di *Maarten Sap et al. 2014* [129] descritto negli aspetti fondamentali in sezione §3.2.1.3.
Nel nostro lavoro x_f viene rappresentata da tutte frequenze relative, come espresso in formula 3.6; di conseguenza abbiamo potuto applicare la formula di classificazione del genere 3.7 con l'ausilio del modello *emnlp14gender* a ciascuna *BOWs*, ottenendo così le formule 4.2 e 4.3 :

$$y` = usage`_{user} + w_0, \tag{4.2}$$

$$usage`_{user} = \sum_{word \in BOWs_{user}} w_{word} * \frac{freq(word, BOWs_{user})}{freq(*, BOWs_{user})}. \tag{4.3}$$

Con la formula 4.2 siamo riusciti a classificare gli utenti in base al genere (se $y` >= 0$ femmina; altrimenti $y` < 0$ maschio).

Oltre all'individuazione del genere di ogni utente, abbiamo anche provveduto alla valutazione del supporto tra le parole di *emnlp14gender* e le *BOWs* per ogni utente e in media, formule 4.4 e 4.5; in modo da valutare in fase di accuratezza dei risultati ottenuti la quota di modello che ha contribuito a classificare il genere di un utente e così stimare l'affidabilità della classificazione.

$$support_{gender}(user) = \frac{\sum_{word` \in BOWs_{user}} \sum_{word \in emnlp14gender} I(word`, word)}{|BOWs_{user}|}, \tag{4.4}$$

con $I = \begin{cases} 1 & \text{se } word\hat{} = word \\ 0 & \text{altrimenti.} \end{cases}$

$$mean_{support\_gender}(User) = \frac{\sum_{user \in User} support_{gender}(user)}{|User|}. \tag{4.5}$$

4.2.2 Risultati

In Figura 4.1 è presentata la classificazione di genere che abbiamo ottenuto.
Il numero di utenti totali che abbiamo ritenuto validi, ovvero con almeno una parola nella *Bag of words*, sono stati 5'189 (su 5'191 persone fisiche e aziende, si veda sezione §3.1.1.1), di cui 1'612 sono risultati essere femmine e 3'577 maschi.

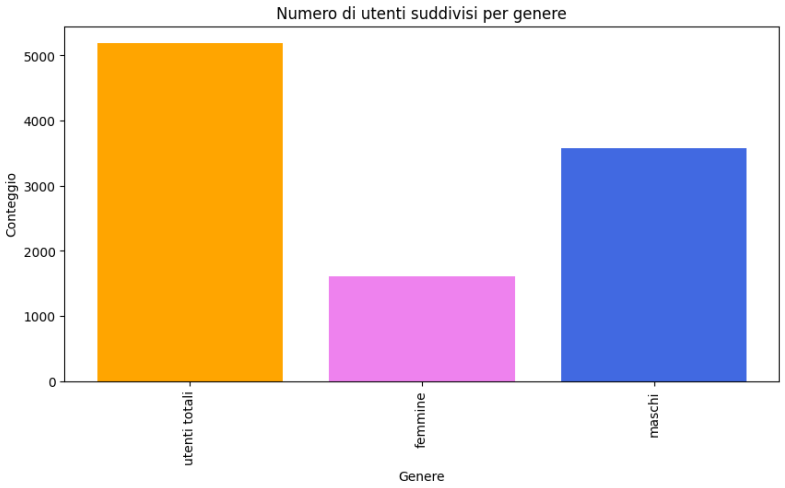


Figura 4.1: *Classificazione di genere - Twitter Occupation Dataset.*

4.2.3 Accuratezza

Per verificare la correttezza della classificazione conseguita, abbiamo considerato un campione di 100 utenti casuali di *Twitter Occupation Dataset*, abbiamo estratto per ciascuno almeno un *tweet id* e ricercato su *Twitter* il post. Un tweet riporta, accompagnato al proprio testo, anche il nome del profilo *Twitter* che l'ha pubblicato, per cui da quest'ultimo siamo riusciti a risalire all'utente del tweet e con la visita al suo profilo *Twitter* siamo stati in grado di individuarne il genere (da bibliografia e dati anagrafici pubblicati dall'utente stesso); in conclusione il genere individuato è stato confrontato con la nostra classificazione.
L'accuratezza che abbiamo riscontrato è mostrata in Figura 4.2, e non considera le persone non fisiche presenti in *Twitter Occupation Dataset*, perchè non è sensato considerarne il genere; per cui il calcolo è su una base di 89 utenti.
Tabella 4.2 mostra la matrice di confusione [133], che si presta molto bene nello stabilire la performance di un algoritmo di classificazione.

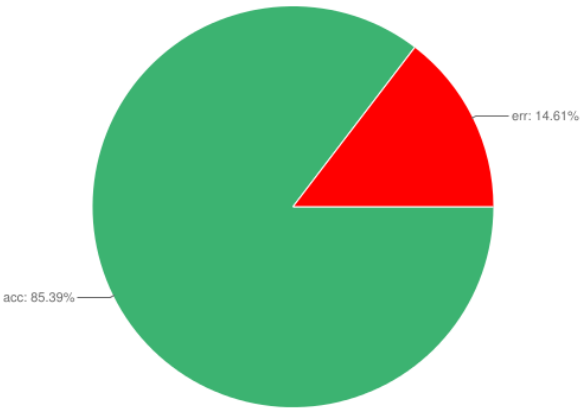


Figura 4.2: *Accuratezza nella classificazione del genere - Twitter Occupation Dataset.*
13 classificazioni di genere errate su 89 (persone fisiche): accuratezza 0.8539.

		predetta	
		femmine	*maschi*
effettiva	*femmine*	29	11
	maschi	2	47

Tabella 4.2: *Matrice di confusione - 89 utenti (persone fisiche) di Twitter Occupation Dataset.*
29 femmine classificate correttamente e 11 assegnate alla classe maschile, 47 maschi classificati correttamente e 2 assegnati alla classe femminile.

Il lavoro [129] presenta un'accuratezza nella classificazione del genere di 0.889 (*corpus* e *test set* su *Twitter*); conseguentemente la nostra classificazione di *Twitter Occupation Dataset* (accuratezza di 0.8539) risulta sufficientemente in linea con l'accuratezza del *test set* di [129], in quanto associamo lo scarto di solo -3.51% della nostra classificazione alla natura del campione casuale che abbiamo estratto per l'analisi.
Per quanto riguarda la matrice di confusione i falsi riscontati (predetta femmine, effettiva maschi e predetta maschi, effettiva femmine) li riconduciamo al supporto del modello nello specifico utente, formula 4.4, per la motivazione riportata di seguito.
Il supporto medio calcolato su tutti gli utenti, formula 4.5, è stato di 0.609 cioè in media il 61% delle parole nelle *BOWs* degli utenti di *Twitter Occupation Dataset* hanno trovato un'associazione nel modello *empln14gender*; un valore discreto che conferma l'accuratezza da noi ottenuta (0.8539). Tuttavia in *BOWs* con un supporto, formula 4.4, risultato molto al di sotto di tale soglia (tipicamente inferiore a 0.50) o molto al di sopra (oltre 0.70) si sono verificate situazioni di errata classificazione (predetta femmine invece di maschi e viceversa); segnale che l'utilizzo scarso o abbondante di un lessico, che può dipendere anche dall'utente stesso e non dai metodi utilizzati, può causare errori di classificazione. Non è stato possibile attribuire tali errori nè al modello nè alla procedura che abbiamo utilizzato per la classificazione del genere, come si può osservare dalle *words clouds* di genere presentate in sezione §4.4.2; inoltre

non abbiamo individuato alcuna motivazione fondata a supporto di un maggior errore di classificazione femminile, come descritto in merito alla ragionevolezza dei modelli e degli approcci utilizzati in sezione §4.4.4; il quale può essere dipeso, ancora una volta, dal campione casuale che abbiamo estratto per verificare la correttezza della classificazione.

4.3 Personalità

Per la predizione dei tratti di personalità *Big Five* abbiamo deciso di istanziare il modello *Word and phrase correlations* [131] (si veda sezione §3.2.1.2) il quale si compone di cinque sottomodelli, uno per ciascun tratto di personalità:

* **A:** *agreeableness* (amichevolezza);
* **C:** *conscientiousness* (coscienziosità);
* **E:** *extraversion* (estroversione);
* **N:** *neuroticism/ emotional stability* (nevroticismo/stabilità emotiva);
* **O:** *openness to experience* (apertura alle esperienze).

Ognuno dei sottomodelli è costituito dalle prime 100 frasi, composte al massimo da 3-grammi, che [131] ha individuato, con l'utilizzo della tecnica di *Differential Language Analysis* (si veda sezione §3.2.1.2), come appartenenti a ogni polarità dei tratti di personalità.
Ciascun sottomodello, Tabella 4.3, si compone dei parametri frase/parola, *correlation* e *p-value*; con correlazione positiva quando la frase/parola appartiene a un alto tratto invece negativa quando a un basso tratto. Inoltre per selezionare le prime 100 frasi, più descrittive per ciascun tratto, gli autori [131] hanno scelto di valutare significative le correlazioni con maggior valore assoluto e p-value inferiore a 4×10^{-9}.
Come per *emnlp14gender*, anche per il modello *Word and phrase correlations* le frasi/parole sono state estratte da *Facebook* da utenti che hanno utilizzato l'applicazione *myPersonality*.

	correlation	**p-value**
birthday wishes !	0.036052	1.17E-17
at church	0.036221	2.86E-18
joy	0.036244	4.74E-18
grateful	0.036311	1.21E-17
hope everyone	0.036494	5.5E-18
so thankful	0.03655	2.73E-18
had an amazing	0.036551	2.1E-18
beautiful day !	0.036661	2.36E-18
happy	0.036828	1.5E-18
much fun	0.036835	2.13E-18
of christ	0.03684	5.91E-19

Tabella 4.3: *Modello Word and phrase correlations, alta amichevolezza.*

Il modello è stato impiegato come riferimento da cui estrarre le correlazioni da assegnare alle parole delle *Bag of words* e applicare con queste la formula di *word-count approach*

(si veda sezione §3.2.1.1). Tale procedura ci ha consentito di predire la personalità di ciascuna *BOWs* di *Twitter Occupation Dataset* per i cinque tratti *Big Five*.

4.3.1 Metodologia

Per predire la personalità degli utenti di *Twitter Occupation Dataset* abbiamo fatto riferimento all'approccio a vocabolario chiuso di *Schwartz et al. 2013*; descritto negli aspetti fondamentali in sezione §3.2.1.1.
Essendo che il nostro scopo era solamente di individuare quali erano i tratti di personalità *Big Five* degli utenti, ci siamo limitati a riutilizzare la formula di *word-count approach* 3.1, con categorie i tratti di personalità e sostituendo la frequenza con la correlazione delle frasi/parole espresse in *Word and phrase correlations* quando presenti anche nella *Bag of words* di un utente (formula 4.6, $correlation(phrase, BOWs_{user})$), tralasciando l'utilizzo della regressione ai minimi quadrati ordinari. Questo ha fatto si che siamo riusciti a predire la correlazione normalizzata di ogni tratto in ciascuna *Bag of words* degli utenti (formula 4.6, $\overline{correlation}(trait|user)$; $\forall trait \in BigFive$).
Un aspetto che abbiamo dovuto affrontare, in formula 3.1, è stato come individuare delle corrispondenze concrete tra ciascun tratto di personalità e le *BOWs* in quanto i sottomodelli sono composti da n-grammi, ovvero frasi, invece nelle *BOWs* si hanno singole parole. Per risolvere tale situazione abbiamo considerato una frase di un sottomodello appartenere alla *BOWs* di un utente quando tutte le parole della frase $((\forall word \in phrase) \in trait)$ erano presenti all'interno della *BOWs* ($word \in BOWs_{user}$), senza tenere conto dell'ordine delle parole che costituivano la frase (formula 4.6, $(\forall word \in phrase) \in trait$; $word \in BOWs_{user}$). La scelta appena espressa ci ha fatto perdere in accuratezza della predizione dei tratti di personalità, ma l'abbiamo considerata una perdita accettabile forzata dalla natura dei dati e del modello che abbiamo utilizzato.

$$\overline{correlation}(trait|user) = \frac{\sum_{\substack{(\forall word \in phrase) \in trait \\ word \in BOWs_{user}}} correlation(phrase, BOWs_{user})}{n}, \tag{4.6}$$

$\forall trait \in BigFive$ e $BigFive = \{A, C, E, N, O\}$,

$$n = |\sum_{\substack{(\forall word \in phrase) \in trait \\ word \in BOWs_{user}}} correlation(phrase, BOWs_{user})|. \tag{4.7}$$

In formula 4.6 il nominatore rappresenta la somma delle correlazioni delle frasi di uno stesso tratto che si trovano anche all'interno della *BOWs* dell'*user*; il denominatore (formula 4.7) rappresenta invece la cardinalità delle frasi, ovvero il numero di correlazioni del modello, che hanno trovato corrispondenza nella *BOWs*. Il rapporto tra questi due fattori ci ha permesso di ottenere per ciascun tratto la correlazione normalizzata effettiva dell'utente.

Con l'uso della formula 4.6 abbiamo predetto per ogni utente di *Twitter Occupation Dataset* il coefficiente di correlazione di ogni tratto di personalità *Big Five*. Inoltre abbiamo calcolato alcune metriche di predizione per ogni utente:

* Il supporto a ogni tratto, formula 4.8:

$$support_{BigFive}(user, trait) = \frac{\sum_{phrase \in trait} \sum_{\forall word \in phrase} \sum_{word` \in BOWs_{user}} I(word`, word)}{|BOWs_{user}|}, \tag{4.8}$$

$$\text{con} \quad I = \begin{cases} 1 & \text{se } word` = word \\ 0 & \text{altrimenti.} \end{cases}$$

* La correlazione media dei cinque tratti, formula 4.9:

$$mean_{correlation}(user) = \frac{\sum_{\forall trait \in BigFive} \overline{correlation}(trait|user)}{|BigFive|}. \tag{4.9}$$

* Il supporto considerando tutti i tratti, formula 4.10:

$$mean_{support\_BigFive}(user) = \frac{\sum_{\forall trait \in BigFive} support_{BigFive}(user, trait)}{|BigFive|}. \tag{4.10}$$

* Se ve ne sono, i tratti alti o bassi per l'utente (si veda sezione §4.3.2).

4.3.2 Polarità dei tratti di personalità

Le informazioni su ciascuna polarità dei tratti di personalità sono state ricavate da *Margaret L. Kern et al. 2014* [76] e riportate in Tabella 4.4:

tratto	alto	basso
Amichevolezza	[0.032, 0.059]	[-0.123, -0.034]
Coscienziosità	[0.035, 0.069]	[-0.105, -0.039]
Estroversione	[0.059, 0.111]	[-0.089, -0.036]
Stabilità emotiva	[0.023, 0.047]	[-0.086, -0.042]
Apertura	[0.072, 0.124]	[-0.090, -0.039]

Tabella 4.4: *Intervallo di correlazione di ciascuna polarità dei tratti di personalità [76].*

4.3.2.1 Definizione delle correlazioni sui tratti

La stabilità emotiva è l'opposto del nevroticismo, per cui per fare riferimento al nevroticismo, presente in *Word and phrase correlations*, ci è stato sufficiente capovolgere il tratto di personalità di Tabella 4.4. Inoltre essendo che la massima correlazione del modello è stata calcolata sulla base della massima correlazione ottenuta dalle frasi e per non limitare i risultati del nostro lavoro, abbiamo esteso gli estremi superiori per gli altri tratti e quelli inferiori per i bassi tratti in modo da considerare anche correlazioni con un valore maggiore dell'estremo indicato nel modello.

tratto	alto	basso
Amichevolezza	[0.032, 1]	[-1, -0.034]
Coscienziosità	[0.035, 1]	[-1, -0.039]
Estroversione	[0.059, 1]	[-1, -0.036]
Nevroticismo	[-1, -0.042]	[0.023, 1]
Apertura	[0.072, 1]	[-1, -0.039]

Tabella 4.5: *Intervallo di correlazione di ciascuna polarità dei tratti di personalità con estensione degli estremi.*

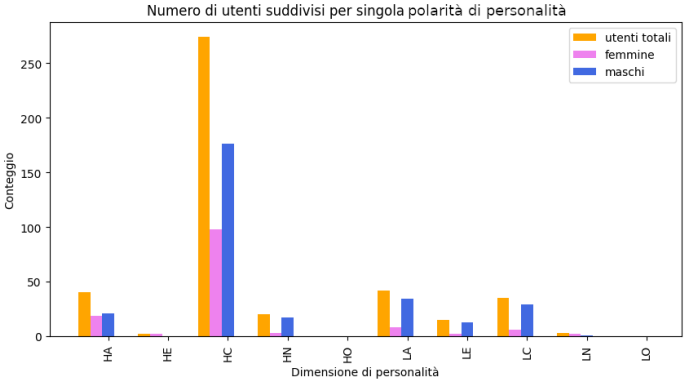
4.3.2.2 Visualizzazione

Per avere maggiore chiarezza nella visualizzazione dei nostri risultati e solo a fini interpretativi delle correlazioni, abbiamo deciso di capovolgere il segno di alto/basso nevroticismo rispetto al tratto originario, Tabella 4.6.

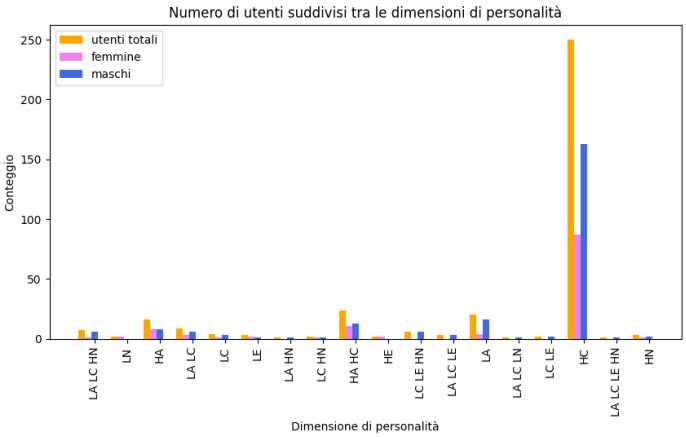
tratto	alto	basso
Amichevolezza	[0.032, 1]	[-1, -0.034]
Coscienziosità	[0.035, 1]	[-1, -0.039]
Estroversione	[0.059, 1]	[-1, -0.036]
Nevroticismo*	[0.042, 1]	[-1, -0.023]
Apertura	[0.072, 1]	[-1, -0.039]

Tabella 4.6: *Interpretazione correlazioni di ciascuna polarità dei tratti di personalità.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.

4.3.3 Risultati



(a) *per singola polarità*



(b) *dimensioni di personalità rilevate*

Figura 4.3: *Predizione delle dimensioni di personalità - Twitter Occupation Dataset.* H: alto tratto; L: basso tratto - A: amichevolezza; C: coscienziosità; E: estroversione; N: nevroticismo; O: apertura alle esperienze.

Le correlazioni, così come definite in Tabella 4.5, sono state applicate a ciascuna personalità predetta, in modo da ottenere per ogni utente l'indicazione della presenza di alti e/o bassi tratti di personalità. In Figura 4.3 è presentata la predizione delle dimensioni di personalità *Big Five* che abbiamo ottenuto.
La polarità che spicca maggiormente è l'alta coscienziosità (*HC*) Figura 4.3a, con 250 utenti su 5'189 che presentano unicamente il tratto e 24 che la vedono accoppiata all'alta amichevolezza (*HA*) Figura 4.3b. Abbiamo anche evidenziato le dimensioni partizionate per genere e ancora una volta è stato *HC* a prevalere Figura 4.3a sia nelle femmine (*HC*:87; *HA HC*:11 Figura 4.3b) che nei maschi (*HC*:163; *HA HC*:13 Figura 4.3b).
Tuttavia l'analisi delle dimensioni non ci ha fornito alcuna informazione attendibile, in quanto ha coinvolto solo il 6.86% degli utenti di *Twitter Occupation Dataset*, un numero troppo esiguo per una tendenza chiara e inconfutabile. Una strategia che invece ha prodotto risultati interessanti è stata l'analisi delle correlazioni medie calcolate su

tutti gli utenti. Difatti considerare la totalità delle correlazioni, riportata in Tabella 4.7, e non esclusivamente un loro sottoinsieme, ci ha permesso di evidenziare che gli utenti della nostra analisi detengono concretamente un discreto livello di coscienziosità (correlazione di coscienziosità media di 0.023; su un alto tratto compreso tra [0.035, 1]).

In Tabella 4.7 sono riportate le correlazioni medie individuate per ogni tratto che abbiamo predetto con la cardinalità di ogni polarità.

tratto	**correlazioni**	**alto**	**basso**
Amichevolezza	0.004	40	42
Coscienziosità	0.023	274	35
Estroversione	0.040	2	15
Nevroticismo*	0.013	20	3
Apertura	0.019	0	0
media	**0.020**	**356/5189 - 6.86%**	

Tabella 4.7: *Correlazioni medie e cardinalità delle dimensioni per ciascun tratto di personalità - Twitter Occupation Dataset.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
Le correlazioni sono state calcolate con $\overline{correlation}(trait|user)$ (formula 4.6) mediata su tutti gli utenti.

4.3.3.1 Approfondimento

Il livello di coscienziosità degli utenti di *Twitter Occupation Dataset* che abbiamo riscontrato (0.023) ci permette di affermare che tendono a essere dei buoni lavoratori [150, 155].
Questo ci ha fornito degli spunti di riflessione riguardanti la connessione tra coscienziosità [150, 155], motivazione in ambito lavorativo [41] e sovraqualificazione [79]. Questi difatti sono elementi che nel dataset presentano, grazie alla coscienziosità media tendente all'alto tratto, una situazione dei lavoratori di buona/discreta soddisfazione. In quanto la sovraqualificazione, largamente diffusa all'interno degli ambiti occupazionali [79], è in grado di impattare sulla performance e sulla soddisfazione di un lavoratore quando questi ha una mansione inferiore rispetto alla propria qualifica [41], riducendone anche la coscienziosità [155]; situazione che la correlazione da noi calcolata non evidenzia.
Anche se un'associazione tra sovraqualificazione e coscienziosità può essere interessante, è tuttavia svolta considerando l'intero dataset, una visione d'insieme troppo grande per permetterci di cogliere tutte le relazioni che possono esistere tra i dati e dare significatività alle evidenze escludendo che siano solo una somma di casualità.

In sezione §4.5 studiamo alcune di queste relazioni (genere e personalità) in rapporto a specifiche categorie occupazionali (STEM, Informatica e Matematica) che permettono di dare maggiore significato ai nostri risultati.

4.3.4 Accuratezza

Per verificare la correttezza della predizione ottenuta nei diversi tratti e avendo a disposizione il tratto di personalità predetto per ciascun utente, abbiamo ritenuto

adeguato valutare la conservazione delle correlazioni tra i tratti di personalità. Difatti se una di queste fosse risultata errata avrebbe denotato un errore di predizione associato ai tratti coinvolti nella correlazione. A tal fine abbiamo svolto analisi correlazionale con metodo di *Pearson* tra i tratti di personalità *Big Five* (maggiori dettagli in sezione §4.5.3).
In Tabella 4.8 sono riportate le correlazioni che abbiamo ottenuto tra i tratti di personalità degli utenti di *Twitter Occupation Dataset*.

tratto	**A**	**C**	**E**	**N**	**O**
A: Amichevolezza	1				
C: Cosienziosità	**0.704**	1			
E: Estroversione	0.218	0.592	1		
N: Nevroticismo*	**-0.651**	-0.532	-0.253	1	
O: Apertura	-0.024	0.174	0.015	0.203	1

Tabella 4.8: *Correlazioni tra i tratti di personalità Big Five - Twitter Occupation Dataset.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Analizzando i risultati riportati in Tabella 4.8 individuiamo amichevolezza e coscienziosità come la più alta correlazione positiva tra tratti di personalità (0.704); invece amichevolezza e nevroticismo come quella negativa (-0.651). Le forze di queste correlazioni sono state riscontrate anche nel lavoro di *Azeem Akbar, Amara Malik e Nosheen Fatima Warraich 2023* [10].
Essendo *Twitter* nell'ambito di un social network ci aspettiamo tuttavia che nella realtà il nevroticismo abbia una correlazione positiva con l'amichevolezza leggermente superiore (*Yusuke Umegaki e Ayaka Higuch 2022* [145], *Amichai-Hamburger et al. 2002* [158]) ma non sufficiente da far valere la correlazione negativa tra amichevolezza e apertura (-0.024), in quanto anche quest'ultima sospettiamo influenzata dall'ambito di analisi.

In Tabella 4.9 sono riassunti i risultati delle metriche di predizione.

tratto	**supporto**
Amichevolezza	0.0371
Coscienziosità	0.0431
Estroversione	0.0489
Nevroticismo*	0.0412
Apertura	0.0512
media	**0.0443**

Tabella 4.9: *Supporto per ciascun tratto di personalità - Twitter Occupation Dataset.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
Il supporto è stato calcolato con $support_{BigFive}$ (formula 4.8) e la media con $mean_{support\_BigFive}$ (formula 4.10) mediate su tutti gli utenti.

Osservando la valutazione media di supporto a ogni tratto e il supporto medio complessivo (0.0443) di Tabella 4.9, e considerandoli molto bassi rispetto all'intera copertura

del modello, riteniamo che le correlazioni da noi individuate tra i tratti A-C, A-N e confermate da [10] indichino la correttezza della predizione di personalità.
Inoltre per ciascuna polarità dei tratti di personalità abbiamo generato delle *word clouds*, descritte in sezione §4.4.3, per valutare la ragionevolezza dei modelli e degli approcci utilizzati; non rilevando alcun errore di predizione di personalià particolare, per esempio su specifici gruppi di parole, eccetto la mancanza di contesto nelle parole come descritto in sezione §4.4.4.

4.4 *Word clouds*

La costruzione di *word clouds* è una tecnica che facilita l'interpretazione di un'alta numerosità di dati (parole) con una specifica *feature* (come genere e tratti di personalità). Nei lavori [131, 76] ad esempio vengono utilizzate all'interno dell'approccio a vocabolario aperto *Differential Language Analysis* (*DLA*, si veda sezione §3.2.1.2), consentendo ai ricercatori di individuare le caratteristiche linguistiche che influenzano le *features* demografiche e psicologiche.
Abbiamo fatto uso delle *word clouds* per verificare la ragionevolezza dei modelli e degli approcci utilizzati per la classificazione del genere descritta in sezione §4.2 e la predizione della personalità descritta in sezione §4.3. A tal fine sono state costruite le *word clouds* considerando i risultati di genere e personalità ottenuti su tutte le parole di *Twitter Occupation Dataset*.

4.4.1 Metodologia

Abbiamo realizzato delle *word clouds* applicate alle *BOWs* di *Twitter Occupation Dataset* per genere e per polarità dei singoli tratti di personalità *Big Five*. In entrambi i casi abbiamo deciso di seguire degli approcci a vocabolario aperto; per il genere abbiamo utilizzato un approccio a *clustering*; invece per la personalità *DLA*.

Lo strumento scelto per la loro generazione è stata la libreria *wordcloud - PyPI* [109], preferita rispetto a *tools* online come *WordClouds* [166] e *WordArt* [157], perchè in questo modo la creazione delle nuvole di parole è stata interamente gestite da noi che abbiamo così potuto definire la dimensione delle parole e il raggruppamento per argomenti, senza alcuna limitazione di scala e nei caratteri non alfanumerici.
Un aspetto che abbiamo dovuto risolvere per ottenere delle *word clouds* con valore è stato quello di definire delle *stop words* personalizzate; in quanto *Twitter Occupation Dataset* è composto da un *corpus* che anche se filtrato sulla lingua inglese contiene numeri, date, link, parole ambigue o in altre lingue che se considerate avrebbero fatto perdere di significato i nostri risultati. A tal fine abbiamo istanziato, per coerenza con il lavoro [131] le *stop words* della libreria *Mallet*, sulle quali abbiamo incluso ulteriori parole (*custom stop words*) definite da noi.

Per ottenere il valore della parole (*feature value*) delle nostre *word clouds* abbiamo realizzato tre strategie:

* ***Term frequency*** (tf_{word} o x_{word}): con riferimento al lavoro [129] (formula 4.1) ci ha permesso di individuare la *feature value* di ciascuna parola di *Twitter Occupation Dataset*. La *term frequency* l'ha abbiamo impiegata per generare le *word clouds* sul genere.

* ***Correlazioni***: tecnica utilizzata dai lavori [131] e [76]. Abbiamo messo in correlazione la probabilità a posteriori ottenuta sulle parole e argomenti (formule 3.4 e 3.5) con ciascun tratto di personalità predetta degli utenti; con il fine di individuare la *feature value* per le *word clouds* sulle polarità dei tratti di personalità *Big Five*.

* ***Term frequency-inverse document frequency***: tecnica alternativa (formula 4.11) che abbiamo studiato come estensione alla *term frequency*:

$$tfidf_{word}(user, User) = tf_{word}(user) * idf_{word}(user, User), \tag{4.11}$$

$$idf_{word}(user, User) = log(\frac{N}{|user \in User : (word \in BOWs_{user}) \in user|}). \tag{4.12}$$

In formula 4.12 il nominatore N è il numero totale di utenti in *Twitter Occupation Dataset*, il denominatore $|user \in User : (word \in BOWs_{user}) \in user|$ è il numero di utenti che contengono la parola *word* nelle loro *BOWs*.

4.4.2 *Word clouds* di genere

Le *feature value* di ogni parola sono state calcolate come la somma delle frequenze nelle *BOWs* normalizzate per la cardinalità dei generi, formula 4.13. Successivamente ciascuna parola non inclusa nelle *stop words*, è stata assegnata a un singolo genere sulla base del valore maggiore della *feature value*, moltiplicata per un fattore costante $(N * 100)$.

$$normalization_{word}(Gender) = \sum_{\substack{word \in BOWs_{user} \\ (\forall user \in User) \in Gender}} \frac{tf_{word}(user)}{|Gender|}, \tag{4.13}$$

con $(\forall i \in Gender \wedge i \in Female) \vee (\forall i \in Gender \wedge i \in Male)$.

Questo procedimento è stato ripetuto sia per strategia *term frequency* che *term frequency - inverse document frequency* per generare le *word clouds* di genere per le parole di *Twitter Occupation Dataset*.

4.4.2.1 Risultati

Stop words. Le *style words* (come *the, in, has, here, is, had*), i numeri e le date presenti nel modello *emnlp14gender* sono state valutate poco significative per il genere in quanto, utilizzando degli approcci basati sulla frequenza, avrebbero ricevuto un eccessivo peso rispetto alla reale attribuzione al genere. Per questo e con il fine di evitare situazioni di incoerenza nell'analisi testuale, criticate in [58] (si veda sezione §3.2.1.2), abbiamo deciso di includerle nelle *custom stop words* [34].

Term frequency - inverse document frequency. Confrontando i risultati di genere ottenuti dagli approcci *term frequency* e *term frequency-inverse document frequency* su tutte le parole di *Twitter Occupation Dataset*, Figura 4.4, ci accorgiamo che le *word clouds* con *tfidf* danno eccessivo valore a parole che poco contraddistinguono il genere (nel caso femminile *mindy, bogota, nicole, gw, tp* - Figura 4.4a) rispetto ad

altre che invece *tf* mette ben in evidenza (come *love, lovely, baby, excited* - Figura 4.4b) [131].

(a) *tfidf*

(b) *tf*

Figura 4.4: *Femmine word clouds - Twitter Occupation Dataset.*
Le 100 parole con *term frequency-inverse document frequency* e *term frequency* maggiore per il genere femminile generate dai risultati di genere ottenuti su tutte le parole di *Twitter Occupation Dataset*.

Term frequency. Di seguito presentiamo le nostre *word clouds* di genere generate dai risultati di genere ottenuti su tutte le parole di *Twitter Occupation Dataset* con approccio *term frequency*.

Femmine *word cloud.* Figura 4.5 mostra come parole tipicamente femminili i termini che esprimono emozioni (*love, excited, lovely, amazing, gorgeous*) [131, 34]; parentela (*mom, child, kids, children, parents*) [131, 34]; abbreviazioni (*xx, lol, xxx,*

xoxo, omg, xxxx) [131, 34] e suoni (*haha, ahhh, hmm, um, ugh*) [34]. Dunque, per quanto affermato, il genere femminile all'interno dei social networks è propenso a esternalizzare i propri sentimenti.

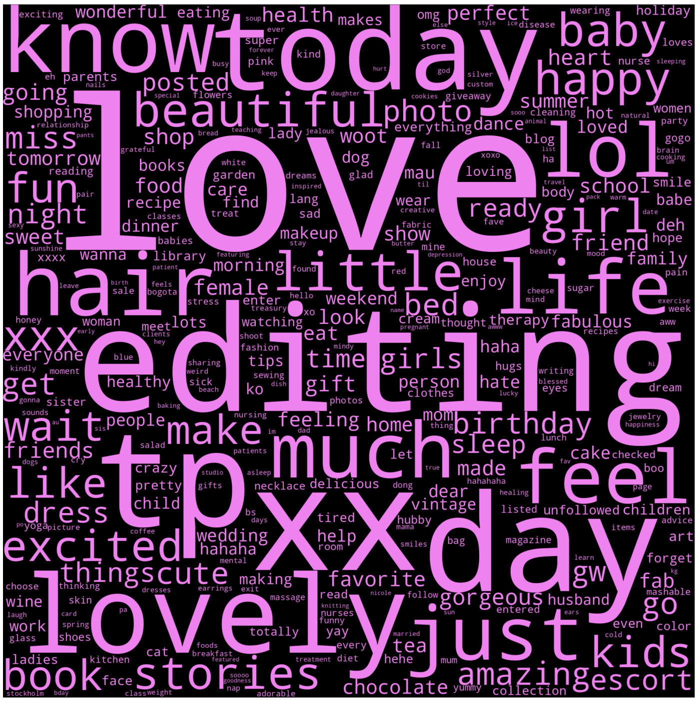


Figura 4.5: *Femmine word clouds - Twitter Occupation Dataset.*
Le 2'500 parole con *term frequency* maggiore per il genere femminile generate dai risultati di genere ottenuti su tutte le parole di *Twitter Occupation Dataset.*

Una particolarità che osserviamo nel nostro lavoro è l'utilizzo della parola *escort*. Da ricerche precedenti come *H. Andrew Schwartz et al. 2013* [131], *Margaret L. Kern et al. 2014* [76], e *Bamman David, Jacob Eisenstein e Schnoebelen Tyler 2014* [34] questo non è un termine che caratterizza il genere femminile; tuttavia nella nostra analisi appare molto usato come conseguenza della natura del nostro dataset. Infatti ricercando su *Twitter* alcuni *tweet id* di *Twitter Occupation Dataset* ci siamo accorti che l'utilizzo di *escort* è da ricondurre a siti di compagnia (che vengono erroneamente dal nostro classificatore etichettati come femmine) e da utenti avvocate e scrittrici impegnate in temi sociali e di difesa della donna. L'unione delle diverse origini ha come effetto l'aumento della *term frequency* della parola *escort*.

Maschi *word cloud*. Figura 4.6 mostra come parole tipicamente maschili i termini che indicano relazioni d'amicizia (*mate, bro, bros, wife*) [34] e parolacce [131, 34]. Inoltre appare evidente un'interesse tecnologico (*game, website, video, app, games, mobile, media, web, account, digital, software, facebook*) [34, 131]; sportivo (*team, football, run, bike, ride, golf*) [131] e politico-finanziario (*business, tax, money, government*) [131]. Dunque, per quanto appena affermato, il genere maschile all'interno dei social networks presenta un maggior interesse verso il mondo materiale.

Figura 4.6: *Maschi word clouds - Twitter Occupation Dataset.*
Le 2'500 parole con *term frequency* maggiore per il genere maschile generate dai risultati di genere ottenuti su tutte le parole di *Twitter Occupation Dataset*.

4.4.3 *Word clouds* di personalità

Per la generazione delle *word clouds* di personalità abbiamo seguito gli aspetti descritti in sezione §3.2.1.2 [131, 76], con alcune modifiche derivanti dalla natura dei nostri dati:

* **Parole e frasi**: essendo in possesso solo di singole parole, non abbiamo potuto svolgere il calcolo di *pmi* (formula 3.2); inoltre noi parliamo di *user* e non di *subject*, e di BOW_{user} e non di $vocab(user)$. Dunque per calcolare la *feature value* delle parole abbiamo modificato le formule 3.3 e 3.4 come segue:

$$p(word|user) = \frac{freq(word, BOWs_{user})}{\sum_{word' \in BOWs_{user}} freq(word', BOWs_{user})}, \tag{4.14}$$

$$p_{ans}(word|user) = 2\sqrt{p(word|user) + \frac{3}{8}}. \tag{4.15}$$

* **Argomenti**: abbiamo impiegato la stessa tecnica *LDA* di *Mallet package*. Con l'output di *LDA* siamo riusciti a realizzare formula 3.5 come segue:

$$p(topic|user) = \sum_{word \in (BOWs_{user}, topic)} p(topic|word) * p_{ans}(word|user), \tag{4.16}$$

$$p(topic|word) = \frac{weight_{LDA}(word, topic)}{\sum_{\forall topic \in word} weight_{LDA}(word, topic)}. \tag{4.17}$$

In formula 4.17 il numeratore rappresenta il peso maggiore del *topic* assegnato alla parola (*majority topic*); invece il denominatore rappresenta la somma di tutti i pesi dei *topics* a cui la parola è stata assegnata dalla procedura.

* **Analisi correlazionale:** abbiamo lavorato con le matrici di correlazione di *Python* e metodo di *Pearson* descritti in sezione §4.5.3.
  Per seguire il più possibile quanto svolto in [131, 76], abbiamo calcolato la correlazione di ciascuna parola e argomento in associazione a ciascun tratto di personalità *Big Five*; confrontando la probabilità a posteriori di *topic* e parola, calcolata su ogni utente, con le polarità dei tratti di personalità.
  Successivamente, con riferimento a quanto riportato in Tabella 4.5, abbiamo selezionato per ciascun alto e basso tratto solo i *topics* con una correlazione inclusa nell'intervallo definito, e per questi solo le parole anch'esse con correlazione inclusa nell'intervallo.
  Una volta ottenute le parole dei *topics* all'interno dei tratti, abbiamo moltiplicato la loro correlazione, come fatto per genere, per un fattore costante.

* **Visualizzazione:** abbiamo generato le *word clouds* per ciascuna polarità dei tratti, con raggruppamento delle parole di *Twitter Occupation Dataset* non incluse nelle *stop words*, per *topics*.

4.4.3.1 *Latent Dirichlet Allocation*

Nella tecnica di *Latent Dirichlet Allocation* noi abbiamo considerato come raccolta di documenti le 200 parole più frequenti delle *BOWs* di ciascun utente e qui vi abbiamo applicato la fase di *topic modeling*.
I passaggi eseguiti sono stati i seguenti:

1. Importazione delle *BOWs* nel formato della libreria *Mallet* (formato *MALLET*) e sul quale lavora la stessa, mantenendo l'ordine della sequenza originale (7'487 parole per 5'189 utenti);

2. Il numero di *topics* sul quale abbiamo costruito il modello è 20. In *H. Andrew Schwartz et al. 2013* [131] per 700 milioni di parole e 75'000 volontari sono stati istanziati 2'000 *topics* (circa 350'000 parole a singolo *topic*); per cui abbiamo valutato per il nostro lavoro 20 una base più che sufficiente per creare raggruppamenti significativi ($17'487/20 = 874$ parole circa a singolo *topic*).

3. Il parametro *alpha* necessario per il calcolo dei pesi delle parole del modello è stato fissato seguendo [131] a 0.30. Il valore di default della libreria è 5.0, la riduzione di questo fa si che vengano generati pochi argomenti per *BOWs*; scelta auspicabile per la trattazione di post.

4.4.3.2 Risultati

Stop words. Nelle *custom stop words* abbiamo deciso di includere quelle parole non *content words* non significative per identificare la personalità (come *till, if, cc, silverback, approaches, crusoe*) [76]. Inoltre abbiamo inserito anche parole di lingua diversa dall'inglese, quando non di uso nel linguaggio comune o con troppe ambiguità.

Correlazione. A seguire le *word clouds* di personalità generate dai risultati di personalità ottenuti su tutte le parole di *Twitter Occupation Dataset* con approccio

correlazionale. Nel dettaglio trattiamo le polarità di alto nevroticismo e bassa estroversione; tratti di personalità *Big Five* che dalle nostre analisi in sezione §4.5.3 presentano correlazioni con le occupazioni STEM.

Alto nevroticismo. Figura 4.7 è incentrata su termini e abbreviazioni tipicamente appartenenti al contesto sessuale (*escorts, escort, voda, sbb, bjb, ovi, zay, mij, seg, devianart, voluptuous*) e omofobo [76] (*transvestite, transsexual, geh, pata, yom, ifc*). Sono presenti anche parolacce/riferimenti religiosi [76] (*isto, osh, samo, efter, satan*) allusioni a droghe e sostanze [76] (*zoot*) e stati d'animo incerti o negativi [76] (*niv, muero, odio*).
Trovano relazione anche nomi di persona (*thurman, mahmood, elisabeth, lizzy, wilhelm, alisha*) e particolari eventi (*tfa* acronimo di *"25th Amendment to the US Constitution"* utilizzato dagli utenti dei social networks per richiedere la rimozione del allora presidente Donald Trump dalla carica) che abbiamo ricondotto a personaggi pubblici; che nel periodo in cui *Twitter Occupation Dataset* è stato generato sono stati coinvolti in scandali sessuali, di *doping* o oggetto di azioni d'odio e razziali.
Osserviamo anche parole che non vengono classicamente identificate con il tratto ma che associamo alle conseguenze di lunatico, preoccupato, irascibile del nevroticismo [70]. Facciamo rientrare in tale contesto i nomi del mondo arabo (*somalis, abdullah, mohammad, arabia, iraq, armenia*) e il termine latino *omnia* [142], cioè qualunque cosa/tutto, raggruppato insieme alle cariche pubbliche *hollande* e *wilhelm*.

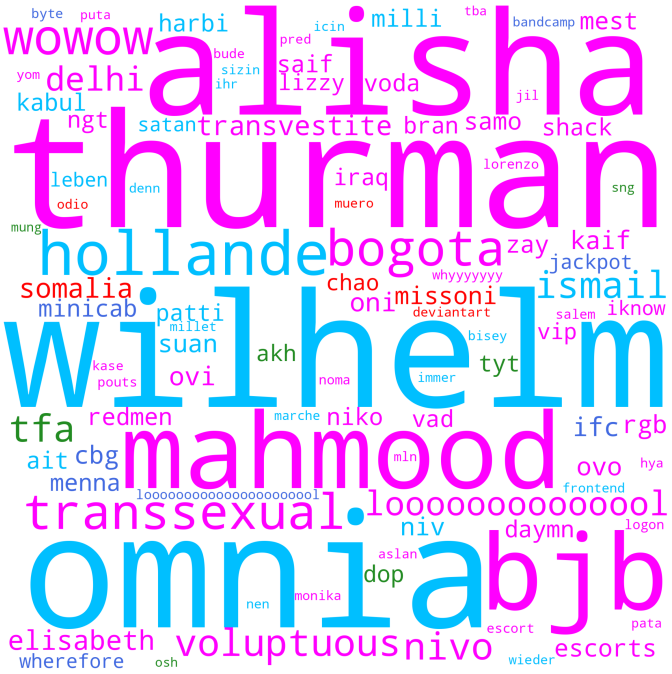


Figura 4.7: *Alto nevroticismo word cloud - Twitter Occupation Dataset.*
Le 100 parole con correlazione maggiore con la polarità dell'alto nevroticismo generate dai risultati di personalità ottenuti su tutte le parole di *Twitter Occupation Dataset*.

Bassa estroversione. Figura 4.8 è molto simile a quanto individuato per l'alto nevroticismo [76]. Compaiono nuovamente termini sessuali e omofobi (*milfs, transsexual, troika, bjb, escorts, transvestite*), le relazioni con i nomi di persona o eventi coinvolti in qualche scandalo (*thurman, mahmood, alisha, tfa*), le parolacce con riferimenti religiosi (*menna, ifc*) e di disgusto (*chyba*). Inoltre la bassa estroversione è caratterizzata da attività isolate [76], questo può dunque spiegare la presenza di parole come *salgado*, noto fotografo e marca di vino brasiliana, e *tapperware*, utensile usato in cucina.
Osserviamo anche riferimenti a feste "fuori dagli schemi" (*sobie, kiedy*), e acronimi legati alla satira e all'enfasi (*loooooool, okok, wowow, pks, baca*). Queste ultime non sono direttamente riconducibili al tratto ma possono essere associate a un'eccessiva timidezza e riservatezza tipiche dell'introversione [125] che possono a nostro avviso sfociare in qualche trasgressione.

Figura 4.8: *Bassa estroversione word cloud - Twitter Occupation Dataset.*
Le 100 parole con correlazione maggiore con la polarità della bassa estroversione generate dai risultati di personalità ottenuti su tutte le parole di *Twitter Occupation Dataset*.

Conclusioni. L'alto nevroticismo/bassa stabilità emotiva e la bassa estroversione sono molto simili, eccetto che l'alto nevroticismo è maggiormente riconducibile a stati d'animo negativi e ansiosi [76] (*odio, muero, niv*).
Per quanto concerne le altre polarità (dettagli in Figura A.5, Appendice A.2):

- Per i bassi tratti, questi rimangono in linea con quanto individuato da alto nevroticismo e bassa estroversione [76]; questo è sottolineato anche dall'uniformità dei *topics* individuati. Evidenziamo inoltre come la bassa apertura si correli con

termini vinicoli (*malbec, vineyards, salgado, santino, missoni*), che ricolleghiamo alla bassa sofisticazione culturale e intellettuale del tratto [76].

* Per gli alti tratti, individuiamo in alta amichevolezza e coscienziosità, riferimenti al benessere del corpo e interventi (*liposuction, hypnobirthing, vaccinations, osteopath, hypnosis, dentistry, reflexology*) che possono indicare attenzione ai risultati, e la ricerca di rilassamento ed equilibrio [76]. Invece riconosciamo nel basso nevroticismo riferimenti ad abbigliamento, all'innovazione (*pinafore, overcoat, playsuit, milinery, khaki, matalan, chiffon, cobalt, titanium, mosaics, carver, R&B*), alla famiglia e all'interesse per la società (*grandkids, tennant, chums, middleton*) che determinano la ricerca di relazioni sociali positive [76].
  Notiamo anche, in alta amichevolezza e coscienziosità, l'utilizzo di parole (come *fff, rtsp*) che dovrebbero incentivare gli utenti a diventare *followers*, indicando la necessità di contatti dei tratti coinvolti.
  Risultati particolarmente anomali, che si discostano dalla letteratura esistente sui tratti di personalità *Big Five*, sono stati rilevati nell'alta apertura che vede la presenza di parole come *abuser* e *knitter*, le quali non hanno nulla a che fare con le frasi sociali che invece ci saremmo aspettati nel tratto. Sono state rilevate incompletezze nella coscieziosità che presenta un'attenzione al benessere fisico, ma nessun riferimento a scuola e lavoro tipici del tratto, nell'amichevolezza con alcuna associazione significativa a termini famigliari e religiosi, e nella non presenza di *topics* e parole correlate con il tratto di alta estroversione.
  Tali discrepanze sono motivate dalla natura di *Twitter Occupation Dataset*, che ci ha permesso esclusivamente di correlare con le polarità dei tratti le *BOWs* degli utenti, senza nessun contesto (come invece accade nei lavori di *Schwartz et al. 2013* [131] e *Margaret L. Kern et al. 2014* [76]). Probabilmente questo non è sufficiente a mettere in luce pienamente le parole che contraddistinguono un tratto di personalità *Big Five*.

4.4.4 Ragionevolezza dei modelli e degli approcci utilizzati

Le evidenze supportate dalla letteratura nelle *words clouds* generate (si veda sezioni §4.4.2 e §4.4.3) indicano una buona qualità dei modelli e degli approcci utilizzati per la classificazione del genere descritta in sezione §4.2 e la predizione della personalità descritta in sezione §4.3. Questo soprattutto per quanto riguarda il genere in linea con l'accuratezza della classificazione (0.8539, si veda sezione §4.2.3) e senza alcuna anomalia che indichi nei nostri risultati la tendenza a classificare erroneamente di un genere anzichè dell'altro. Invece per quanto riguarda la personalità, a causa della mancanza di contesto nelle parole (si veda sezione §4.4.3.2), tale valutazione è più complessa; tuttavia anche in questo caso riteniamo di aver ottenuto dei risultati soddisfacenti e a supporto di quanto già dichiarato in sezione §4.3.4.

4.5 Occupazioni STEM

La categoria di occupazioni oggetto di questo lavoro ha coinvolto le discipline STEM. Per riuscire a tracciare le occupazioni e gli utenti STEM all'interno di *Twitter Occupation Dataset* rimanendo conformi all'*UK SOC code*, abbiamo impiegato come riferimento il lavoro [40] (si veda sezione §3.1.2). Nel dettaglio abbiamo deciso di svolgere un'analisi occupazionale sulle occupazioni STEM individuate dalla lista ridefinita con l'aggiunta del profilo di statistici e matematici (*Business Research and Administrative Professionals (242)*).

Inoltre abbiamo voluto esaminare anche cosa accade in alcune specifiche discipline STEM, quali l'Informatica e la Matematica. A tal scopo, e per individuare rispettivamente i *CS UK SOC codes* e i *MATH UK SOC codes*, abbiamo utilizzato il sito governativo del Regno Unito [121].

4.5.1 Metodologia

La nostra *pipeline* delle attività durante l'analisi occupazionale è stata la seguente:

1. **Analisi preliminare**: per individuare occupazioni e utenti STEM presenti in *Twitter Occupation Dataset*, e le discipline di maggiore incidenza femminile e maschile;

2. **Analisi correlazionale:** generando matrici di correlazione [86] su cui abbiamo ricercato le correlazioni positive e negative più rilevanti tra tratti di personalità, genere e occupazioni, e indagato le possibili evidenze. In questo modo durante l'analisi correlazionale abbiamo realizzato due studi:

    * Impatto della personalità e genere sulla scelta occupazionale, considerando come variabile indipendente l'occupazione (STEM, CS e MATH) e come variabili dipendenti le correlazioni ottenute durante la predizione di personalità *Big Five* di *Twitter Occupation Dataset*, sulla base del genere;

    * Impatto della personalità sul genere, considerando come variabile indipendente il genere e come variabili dipendenti le correlazioni ottenute durante la predizione di personalità *Big Five* di *Twitter Occupation Dataset*. Questa seconda analisi ci è stata utile per comprendere l'incidenza della scelta occupazionale STEM.

4.5.2 Analisi preliminare

In Tabella 4.10 sono riportate le occupazioni STEM con relativo *SOC code*, le discipline di appartenenza e il numero di utenti totali, femmine (f) e maschi (m) che svolgono ciascuna occupazione.

STEM Job family (occupation *STEM UK SOC* code)	**disciplina**	**utenti**
Information Technology Technicians (313)	STEM, CS	141: 6f, 135m
Quality and Regulatory Professionals (246)	STEM	158: 19f, 139m
Business Research and Administrative Professionals (242)	STEM, MATH	69: 16f, 53m
Engineering Professionals (212)	STEM	123: 30f, 93m
Information Technology and Telecommunications Professionals (213)	STEM, CS	126: 8f, 118m
Science Engineering and Production Technicians (311)	STEM	68: 13f, 55m
Functional Managers and Directors (113)	STEM, CS	177: 30f, 147m
Managers and Proprietors in Other Services (125)	STEM	112: 42f, 70m
Public Services and Other Associate Professionals (356)	STEM, CS	64: 14f, 50m
Natural and Social Science Professionals (211)	STEM, CS	145: 44f, 101m
Electrical and Electronic Trades (524)	STEM, CS	98: 14f, 84m
Conservation and Environment Professionals (214)	STEM	140: 31f, 109m

Tabella 4.10: *Occupazioni STEM, CS e MATH - Twitter Occupation Dataset.*
f indica il numero di utenti femmine; m indica il numero di utenti maschi.

In Figura 4.9 sono riportate le distribuzioni degli utenti per genere e disciplina di appartenenza. STEM comprende 1'421 lavoratori di cui 267 sono femmine e 1'154 sono maschi; CS ne comprende 751 di cui 116 sono femmine e 635 sono maschi; infine MATH con 69 di cui 16 femmine e 53 maschi.

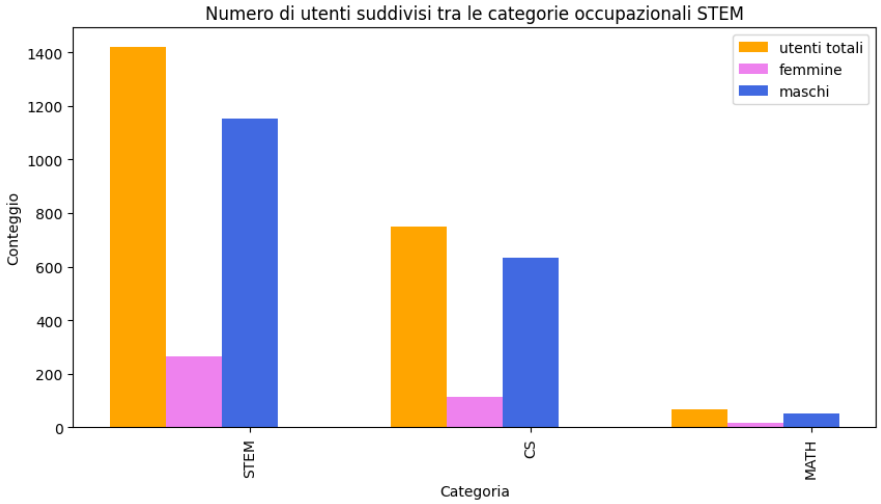


Figura 4.9: *Distribuzione degli utenti STEM tra le discipline - Twitter Occupation Dataset.*

Ispezionando i dati di Tabella 4.10 e Figura 4.9 giungiamo alle seguenti conclusioni:

* **STEM**: le femmine nelle discipline STEM sono nettamente inferiori ai loro colleghi maschi; 18.79% contro 81.21%. In aggiunta anche all'interno delle occupazioni STEM sono emerse preferenze di genere.
  Le femmine preferiscono lavori STEM con un maggior approccio sociale/umano e un'utilità più marcata verso terzi, le più incidenti sono *Natural and Social Science Professionals (211)* e *Managers and Proprietors in Other Services (125)*; inoltre sono meno propense a scegliere professioni davanti al terminale o simili, come *Information Technology and Telecommunications Professionals (213)* e *Information Technology Technicians (313)*. Mentre i maschi preferiscono lavori

più tecnici e davanti al terminale, i più incidenti sono *Information Technology Technicians (313)*, *Information Technology and Telecommunications Professionals (213)* e *Quality and Regulatory Professionals (246)* con alcune eccezioni riguardo le pubbliche relazioni in mansioni manageriali, *Functional Managers and Directors (113)*. Inoltre sono meno propensi a scegliere quelle professioni che richiedono empatia e relazioni sociali, come *Public Services and Other Associate Professionals (356)*.

* **Informatica:** il *gender gap* è ancora più pressante rispetto al caso STEM; 15.45% contro 84.55% (+3.34%). Questo perchè l'Informatica non comprende occupazioni come *Engineering Professionals (212)* (femmine 24.39%) e *Managers and Proprietors in Other Services (125)* (femmine 37.50%) le quali rispetto alle altre occupazioni STEM contengono una sensibile quota di donne.

* **Matematica:** le femmine rimangono in numero inferiore rispetto ai maschi; 23.19% contro 76.81%, ma con un *gender gap* inferiore sia rispetto all'ambiente informatico (−7.74%) che al più generale STEM (−4.4%). Tale aspetto noi lo consideriamo come un indicatore della preferenza delle donne verso carriere matematiche come attuari, economisti e statistici rispetto a quelle informatiche.

In Tabella 4.11 sono presentati i primi 10 lavori STEM, che dalla nostra analisi, risultano maggiormente associati al *gender gap*.

STEM Job family (occupation *STEM UK SOC* code)	**lavori**	**statistiche di genere**
Information Technology Technicians (313)	*Computer games tester, Database administrator, IT technician, Network administrator, Systems administrator, Customer support analyst, Help desk operator, IT support technician, Systems support officer*	4.26% femmine 95.74% maschi su 141 utenti
Quality and Regulatory Professionals (246)	*Planning engineer, Quality assurance engineer, Quality control officer (professional), Quality engineer, Compliance manager, Financial regulator, Patent attorney, Quality assurance manager, Quality manager, Air pollution inspector*	12.03% femmine 87.97% maschi su 158 utenti
Engineering Professionals (212)	*Building engineer, Civil engineer (professional), Highways engineer, Petroleum engineer, Public health engineer, Site engineer, Structural engineer, Aeronautical engineer (professional), Aerospace engineer, Automotive engineer (professional)*	24.39% femmine 75.61% maschi su 123 utenti
Information Technology and Telecommunications Professionals (213)	*Data centre manager, IT manager, IT support manager, Network operations manager (computer services), Service delivery manager, Implementation manager (computing), IT project manager, Programme manager (computing), Project leader (software design), Business analyst (computing)*	6.35% femmine 93.65% maschi su 126 utenti
Functional Managers and Directors (113)	*Investment banker, Treasury manager, Marketing director, Sales director, Bid manager, Purchasing manager, Account director (advertising), Head of public relations, Human resources manager, Personnel manager*	16.95% femmine 83.05% maschi su 177 utenti
Managers and Proprietors in Other Services (125)	*Estate manager, Facilities manager, Landlord (property management), Property manager, Garage director, Garage owner, Manager (repairing: motor vehicles), Hairdressing salon owner, Health and fitness manager, Manager (beauty salon)*	37.50% femmine 62.50% maschi su 112 utenti
Public Services and Other Associate Professionals (356)	*Civil servant (HEO, SEO), Higher executive officer (government), Principle revenue officer (local government), Senior executive officer (government), Employment adviser, Human resources officer, Personnel officer, Recruitment consultant, IT trainer, NVQ assessor*	21.88% femmine 78.12% maschi su 64 utenti
Natural and Social Science Professionals (211)	*Analytical chemist, Chemist, Development chemist, Industrial chemist, Research chemist, Biomedical scientist, Forensic scientist, Horticulturist, Microbiologist, Pathologist*	30.34% femmine 69.66% maschi su 145 utenti

Tabella 4.11: *10 lavori associati alle STEM Job family causa del gender gap - Twitter Occupation Dataset.*

4.5.3 Analisi correlazionale

Ai nostri scopi l'analisi correlazionale, con lo studio dei coefficienti, ci ha permesso di analizzare se e quali tratti di personalità sono legati a specifiche occupazioni.

4.5.3.1 Calcolo dei coefficienti

Per il calcolo dei coefficienti di correlazione abbiamo deciso di valutare le correlazioni di ciascun lavoratore e di non limitarci, visti i risultati emersi in sezione §4.3.3, esclusivamente alle polarità dei tratti.

4.5.3.2 Scelta del coefficiente di correlazione

Esistono diverse tipologie di coefficienti di correlazione, i più comuni sono *Pearson* [20, 77] tipicamente utilizzato con la distribuzione normale di due variabili quantitative; *Rho di Spearman* [35] per qualsiasi distribuzione di due variabili ordinali, d'intervallo o di rapporto; *Tau di Kendall* [46] segue i medesimi criteri di *Spearman*, tuttavia è da considerarsi più sensibile agli errori soprattutto con dati di grandi dimensioni [26].
Per la nostra analisi correlazionale abbiamo scelto di utilizzare il coefficiente di correlazione di *Pearson*. Tale scelta ci è stata dettata:

* dalla distribuzione dei dati, che non è perfettamente normale, ma è "quasi" una normale (Figura 4.10) [132];

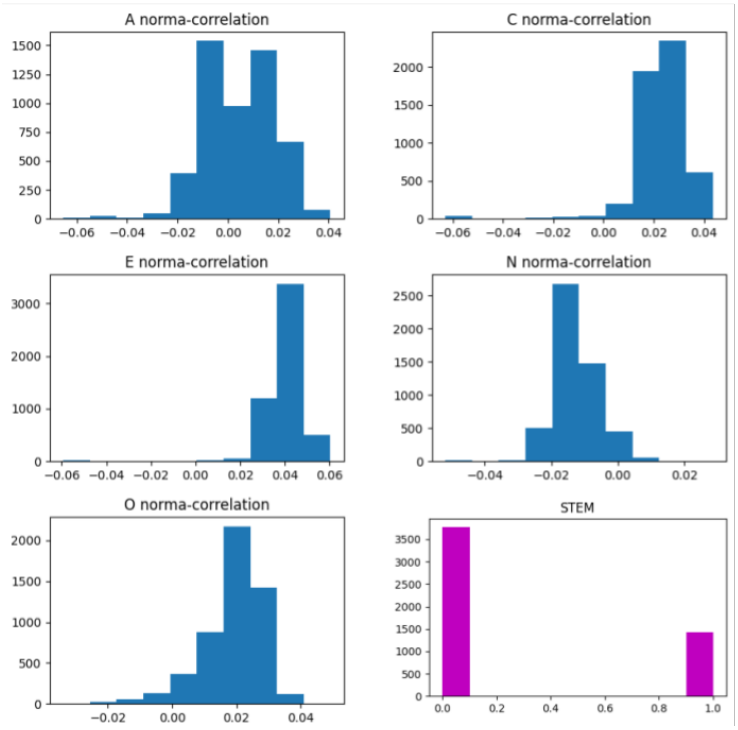


Figura 4.10: *Distribuzione "quasi" normale dei dati in correlazione - Twitter Occupation Dataset.*
Nell'asse delle ascisse sono poste le variabili coinvolte nell'analisi correlazionale, nell'asse delle ordinate le quantità.

* dalla natura dei dati, in quanto abbiamo confrontato tra di loro variabili quantitative (alcuni studi [160, 82] hanno osservato che *Pearson* si adatta bene anche a classi dicotomiche artificiali, come lo sono STEM/NOSTEM, fornendo lo stesso risultato della *Point biserial's correlation* [82, 36]);

* la correlazione di *Pearson* è l'approccio che è stato maggiormente utilizzato in analisi correlazionali simili alla nostra [131, 76, 87].

4.5.3.3 STEM

Utenti STEM	A	C	E	N	O	STEM
A: Amichevolezza	1					
C: Cosienziosità	**0.704**	1				
E: Estroversione	0.218	0.592	1			
N: Nevroticismo*	**-0.651**	-0.531	-0.253	1		
O: Apertura	-0.024	0.174	0.015	0.203	1	
STEM	-0.040	-0.084	**-0.145**	**0.025**	-0.072	1

Tabella 4.12: *Correlazioni tra i tratti di personalità Big Five e STEM - tutti gli utenti STEM.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Femmine STEM	A	C	E	N	O	STEM
A: Amichevolezza	1					
C: Cosienziosità	**0.728**	1				
E: Estroversione	0.176	0.467	1			
N: Nevroticismo*	**-0.653**	-0.503	-0.135	1		
O: Apertura	-0.099	0.123	-0.128	0.274	1	
STEM	**0.028**	-0.032	**-0.075**	-0.007	-0.067	1

Tabella 4.13: *Correlazioni tra i tratti di personalità Big Five e STEM - utenti femmine STEM.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Maschi STEM	A	C	E	N	O	STEM
A: Amichevolezza	1					
C: Cosienziosità	**0.695**	1				
E: Estroversione	0.225	0.628	1			
N: Nevroticismo*	**-0.656**	-0.547	-0.302	1		
O: Apertura	0.001	0.189	0.060	0.173	1	
STEM	-0.048	-0.088	**-0.147**	**0.041**	-0.065	1

Tabella 4.14: *Correlazioni tra i tratti di personalità Big Five e STEM - utenti maschi STEM.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Correlazioni tra i tratti di personalità *Big Five*. Per quanto riguarda le discipline STEM, Informatica (si veda sezione §4.5.3.4) e Matematica (si veda sezione §4.5.3.5), e le correlazioni con genere (si veda sezione §4.5.3.6) valgono le medesime considerazioni fatte in sezione §4.3.4.

Correlazioni con STEM. È stato dimostrato da lavori precedenti che chi intraprende una carriera STEM è significativamente più aperto alle esperienze [28, 27]; di contro risulta meno estroverso [28, 27], meno amichevole [28], meno coscienzioso [27] e con maggior stabilità emotiva [27]. Dai nostri risultati in Tabella 4.12 osserviamo che chi, con riferimento agli utenti di *Twitter Occupation Dataset*, ha intrapreso una carriera STEM ha una correlazione negativa con coscienziosità, amichevolezza ed estroversione [28, 27] tuttavia non riusciamo ad affermare la correlazione positiva con stabilità emotiva e apertura. Questo risultato può essere spiegato dall'utilizzo del social network; difatti [145] ha dimostrato che chi fa uso di *Twitter* ottiene punteggi inferiori in estroversione, coscienziosità e apertura, e più alti per il nevroticismo rispetto a chi non lo utilizza. La minor estroversione e il maggiore nevroticismo presente in chat e social networks è confermato anche da [158]. Non escludiamo in aggiunta la possibilità che l'elevato nevroticismo sia dovuto in parte alla quota non indifferente di utenti CS in STEM (751/1'421, 52.85% circa); il lavoro [127] ha mostrato come già durante gli studi di programmazione, prima di entrare nel mondo del lavoro, gli informatici presentano un alto nevroticismo.

Per quanto concerne le differenze di personalità tra femmine e maschi dalle Tabelle 4.13 e 4.14 ci accorgiamo che le donne STEM sono meno introverse [27], più amichevoli [124] e meno nevrotiche degli uomini. Il nevroticismo per i maschi anzichè per le femmine va dunque contro quanto osservato [87, 124]. Tuttavia la presenza di nevroticismo nelle donne STEM non sembra essere così netta; ne è un esempio in questa direzione [27] da cui risulta che le donne in occupazioni STEM hanno una stabilità emotiva superiore ai colleghi uomini. Inoltre [127] osserva un alto nevroticismo tra le classi di *pair programming* ove il 75% è a presenza maschile.
Un'altra possibile risposta alla correlazione opposta al genere nel nevroticismo è data dallo studio [75], dal quale risulta che le donne che usano *Facebook* sono in parte minoritaria nevrotiche. Ecco che se in CS ci fossero alti tratti di nevroticismo per le femmine [75] comporterebbe che sarebbero restie nell'utilizzo dei social networks, causando una correlazione tra femmine STEM e nevroticismo inferiore della realtà (in quanto il 43.45% delle femmine STEM sono CS, Figura 4.9). Di questo potrebbe essere un indicatore la stabilità emotiva rilevata in MATH per entrambi i generi (-0.037 femmine e -0.031 maschi, Tabelle 4.19 e 4.20) e il nevroticismo solo per i maschi (0.041, Tabelle 4.13 e 4.14) rilevato nelle discipline STEM con maggior incidenza in CS (0.042, Tabelle 4.16 e 4.17). Tuttavia non abbiamo trovato alcun lavoro in letteratura a supporto di tale evidenza.

4.5.3.4 Informatica

Utenti CS	A	C	E	N	O	CS
A: Amichevolezza	1					
C: Cosienziosità	**0.704**	1				
E: Estroversione	0.218	0.592	1			
N: Nevroticismo*	**-0.651**	-0.531	-0.253	1		
O: Apertura	-0.024	0.174	0.015	0.203	1	
CS	-0.045	-0.080	**-0.112**	**0.026**	-0.063	1

Tabella 4.15: *Correlazioni tra i tratti di personalità Big Five e CS - tutti gli utenti CS.* * correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Femmine CS	A	C	E	N	O	CS
A: Amichevolezza	1					
C: Cosienziosità	**0.728**	1				
E: Estroversione	0.176	0.467	1			
N: Nevroticismo*	**-0.653**	-0.503	-0.135	1		
O: Apertura	-0.099	0.123	-0.128	0.274	1	
CS	**0.035**	0.014	-0.020	-0.012	**-0.040**	1

Tabella 4.16: *Correlazioni tra i tratti di personalità Big Five e CS - utenti femmine CS.* * correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Maschi CS	A	C	E	N	O	CS
A: Amichevolezza	1					
C: Cosienziosità	**0.695**	1				
E: Estroversione	0.225	0.628	1			
N: Nevroticismo*	**-0.656**	-0.547	-0.302	1		
O: Apertura	0.001	0.189	0.060	0.173	1	
CS	-0.056	-0.094	**-0.120**	**0.042**	-0.062	1

Tabella 4.17: *Correlazioni tra i tratti di personalità Big Five e CS - utenti maschi CS.* * correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Correlazioni con CS. Le nostre osservazioni rimangono pressochè simili al caso STEM; il che è motivato dal fatto che nel 52.85% circa degli utenti di *Twitter Occupation Dataset* che abbiamo associato a un'occupazione STEM, questa è anche un'occupazione Informatica. Le differenze si rilevano nel nevroticismo, Tabella 4.15, leggermente superiore nel caso CS rispetto al caso STEM (da 0.025 a 0.026) con femmine informatiche che sono più tendenti alla stabilità emotiva (da -0.007 a -0.012, Tabella 4.16) e invece maschi che sono leggermente più nevrotici (da 0.041 a 0.042, Tabella 4.17). La presenza di alto nevroticismo tra i maschi la spieghiamo con [127] che, come già abbiamo spiegato nel caso STEM, riscontra alto nevroticismo in classi di *pair programming* a predominanza maschile.

Osserviamo anche che le femmine tendono sempre alla bassa estroversione (-0.020), ma meno rispetto al caso STEM (-0.075) e questo fa emergere la bassa correlazione con l'apertura, questa ultima dovuta all'uso dei social networks [145, 158]. Per i maschi la situazione è la medesima ma con minore incidenza.

4.5.3.5 Matematica

Utenti MATH	A	C	E	N	O	MATH
A: Amichevolezza	1					
C: Cosienziosità	**0.704**	1				
E: Estroversione	0.218	0.592	1			
N: Nevroticismo*	**-0.651**	-0.531	-0.253	1		
O: Apertura	-0.024	0.174	0.015	0.203	1	
MATH	**0.010**	-0.009	-0.017	**-0.033**	-0.020	1

Tabella 4.18: *Correlazioni tra i tratti di personalità Big Five e MATH - tutti gli utenti MATH.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Femmine MATH	A	C	E	N	O	MATH
A: Amichevolezza	1					
C: Cosienziosità	**0.728**	1				
E: Estroversione	0.176	0.467	1			
N: Nevroticismo*	**-0.653**	-0.503	-0.135	1		
O: Apertura	-0.099	0.123	-0.128	0.274	1	
MATH	**0.019**	-0.010	-0.036	**-0.037**	0.008	1

Tabella 4.19: *Correlazioni tra i tratti di personalità Big Five e MATH - utenti femmine MATH.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Maschi MATH	A	C	E	N	O	MATH
A: Amichevolezza	1					
C: Cosienziosità	**0.695**	1				
E: Estroversione	0.225	0.628	1			
N: Nevroticismo*	**-0.656**	-0.547	-0.302	1		
O: Apertura	0.001	0.189	0.060	0.173	1	
MATH	**0.009**	-0.007	-0.009	**-0.031**	-0.029	1

Tabella 4.20: *Correlazioni tra i tratti di personalità Big Five e MATH - utenti maschi MATH.*
* correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Correlazioni con MATH. Nel caso della Matematica, sia rispetto a STEM che a Informatica, le correlazioni che riscontriamo con i tratti di personalità subiscono

delle modifiche. I matematici hanno una correlazione positiva con amichevolezza [103] e negativa con nevroticismo [103], che diventa basso nevroticismo, Tabella 4.18. Per quanto riguarda il genere, Tabelle 4.19 e 4.20, in entrambi spicca la correlazione media positiva ancora con l'amichevolezza, leggermente superiore per le femmine e la correlazione media negativa con il nevroticismo, nuovamente leggermente superiore per le femmine. Sia l'amichevolezza che il nevroticismo femminile possono essere associati a quanto già discusso nel caso STEM. Osserviamo inoltre che le femmine MATH sono meno estroverse (-0.036) sia della loro controparte maschile (-0.009) che delle femmine CS (-0.020); tuttavia non abbiamo trovato alcun lavoro in letteratura a supporto di tale evidenza.
Essendo che siamo in ambiente *Twitter* ci aspettiamo che il punteggio ottenuto in estroversione e in stabilità emotiva nella realtà siano leggermente superiori [145], soprattutto nel caso maschile [75].

4.5.3.6 Correlazioni con genere

Utenti	A	C	E	N	O	Femmine	Maschi
A: Amichevolezza	1						
C: Cosienziosità	**0.704**	1					
E: Estroversione	0.218	0.592	1				
N: Nevroticismo*	**-0.651**	-0.532	-0.253	1			
O: Apertura	-0.024	0.174	0.015	0.203	1		
Femmine	0.072	0.062	**0.101**	**0.021**	0.048	1	
Maschi	-0.072	-0.062	**-0.101**	**-0.021**	-0.048	-1	1

Tabella 4.21: *Correlazioni tra i tratti di personalità Big Five - utenti maschi e femmine.* * correlazione di segno opposto rispetto a quella calcolata, per coerenza con la polarità del tratto.
correlazione positiva e **correlazione negativa**

Correlazioni con genere. Per quanto riguarda il genere in Tabella 4.21 :

- Le femmine risultano più estroverse, amichevoli, coscienziose e aperte alle esperienze rispetto ai maschi [161];
- Le femmine risultano anche più nevrotiche dei maschi [72, 161].

Considerando che l'ambiente in analisi è quello dei social networks e in particolare di *Twitter* ci aspettiamo che il nevroticismo e l'estroversione per le femmine sia da medio positivo a tendente all'alto tratto [145, 158, 75]; invece per i maschi da medio negativo ad alto tratto per l'estroversione e da medio negativo a basso tratto per il nevroticismo [145, 158].

4.5.4 Conclusioni

In Tabella 4.22 sono riassunti i risultati che abbiamo ottenuto dall'analisi occupazionale. Le evidenze riportano la dicitura *verificata* quando studi precedenti sono giunti alla medesima conclusione; altrimenti si tratta di *osservata*.
Si noti come la scelta di una carriera STEM è guidata da una correlazione negativa con l'estroversione in ambo i sessi [28, 27]; per il nevroticismo invece i nostri risultati non sono stati chiari, sembra però un elemento più caratteristico della programmazione

[127] che del genere (forse maschile, tuttavia mancano studi dettagliati a conferma al riguardo).

Analisi preliminare descritta in sezione §4.5.2	
osservata	Le femmine STEM prediligono carriere che includano un maggior rapporto sociale e umano, con un fine di utilità marcato verso terzi (si veda sezione §4.5.2)
osservata	I maschi STEM prediligono lavori al terminale e posizioni manageriali (si veda sezione §4.5.2)
osservata	Preferenza delle donne verso carriere matematiche come attuari, economisti e statistici rispetto a quelle informatiche (si veda sezione §4.5.2)
Analisi correlazionale descritta in sezione §4.5.3	
verificata	Le femmine sono più nevrotiche degli uomini, ma più aperte alle esperienze, estroverse, amichevoli e coscienziose (si veda sezione §4.5.3.6)
verificata	L'utilizzo dei social networks attrae utenti più nevrotici e introversi. Le femmine tuttavia se nevrotiche preferiscono non farne uso (si veda sezioni §4.5.3.3, §4.5.3.4 e §4.5.3.5)
verificata	Correlazione positiva tra amichevolezza e coscienziosità, negativa tra amichevolezza e nevroticismo (si veda sezioni §4.3.4, §4.5.3.3, §4.5.3.4, §4.5.3.5 e §4.5.3.6)
verificata	I lavoratori STEM hanno una correlazione negativa con l'estroversione (si veda sezioni §4.5.3.3, §4.5.3.4 e §4.5.3.5)
verificata	I matematici hanno una correlazione positiva con l'amichevolezza e negativa con il nevroticismo - stabili emotivamente (si veda sezione §4.5.3.5)
osservata	Il nevroticismo in ambito STEM è causato da occupazioni di programmazione e davanti al terminale (si veda sezioni §4.5.3.3 e §4.5.3.4)
osservata	Le femmine CS sono restie a utilizzare i social networks (si veda sezione §4.5.3.3)
osservata	Le femmine MATH sono meno estroverse dei maschi MATH e delle femmine CS (si veda sezione §4.5.3.5)

Tabella 4.22: *Conclusioni dell'analisi occupazionale - Twitter Occupation Dataset.*

Capitolo 5

Analisi della pubblicità mirata online

In questo capitolo presentiamo l'analisi che abbiamo svolto sulla pubblicità mirata online, e come questa è integrata all'interno di alcuni social networks e motori di ricerca; inoltre spieghiamo come sarebbe possibile svolgere *targeting* psicografico su *Facebook*, e come questo potrebbe aiutare ad alleviare il *gender gap* in ambito STEM.

5.1 Pubblicità mirata online

La prima parte dell'analisi sulla pubblicità mirata online riguarda lo studio di tale tipologia di pubblicità, con lo scopo di comprendere se è possibile realizzare il *targeting* psicografico all'interno dei Media digitali.

5.1.1 Metodologia

I media analizzati sono stati *Facebook*, *YouTube*, *Instagram* e *Google*. In questi ci siamo concentrati, prima di tutto, nel capirne il funzionamento nella pubblicità mirata online e come avviene la protezione dei dati degli utenti; successivamente nelle conclusioni tali aspetti sono stati discussi in rapporto al *targeting* psicografico.

5.1.2 Funzionamento

La pubblicità mirata online, anche conosciuta come *targeted digital advertising* [45], è una forma di pubblicità su *Display*[g] diretta a un pubblico con determinate caratteristiche e sul prodotto che un inserzionista sta cercando di promuovere. Esistono diverse tipologie di pubblicità mirata online sulla base del target che si desidera raggiungere:

* ***Targeting* contestuale:** basato sul contenuto di un sito Web visitato o sulla specifica query di ricerca inserita da un utente in un motore di ricerca;

* ***Targeting* comportamentale:** basato sulle informazioni che gli utenti condividono con una piattaforma Web; ad esempio, post, like, recensioni scritte, cronologia e informazioni tecniche (come produttore del dispositivo, indirizzo IP e fornitore del browser). Il *targeting* psicografico è un estensione del *targeting* comportamentale.

- ***Targeting* segmentato:** basato sulle informazioni di contatto, fornite durante la registrazione dell'utente a un sito Web.

Sono state le tecnologie dell'informazione e della comunicazione che hanno permesso l'estensione del concetto di pubblicità mirata, passando dalla pubblicità tradizionale (su cartelloni, giornali, riviste e canali radiofonici) a quella online/digitale su tutte le tecnologie, Web e mobile.
Nel 2022 la pubblicità mirata online stimata è stata di 441 miliardi [92], di cui la quota maggiore su social media; difatti il 33% [96] è avvenuta con pubblicazione mirata di annunci pubblicitari sotto forma di post testuali, immagini e video su social networks e reti aziendali.

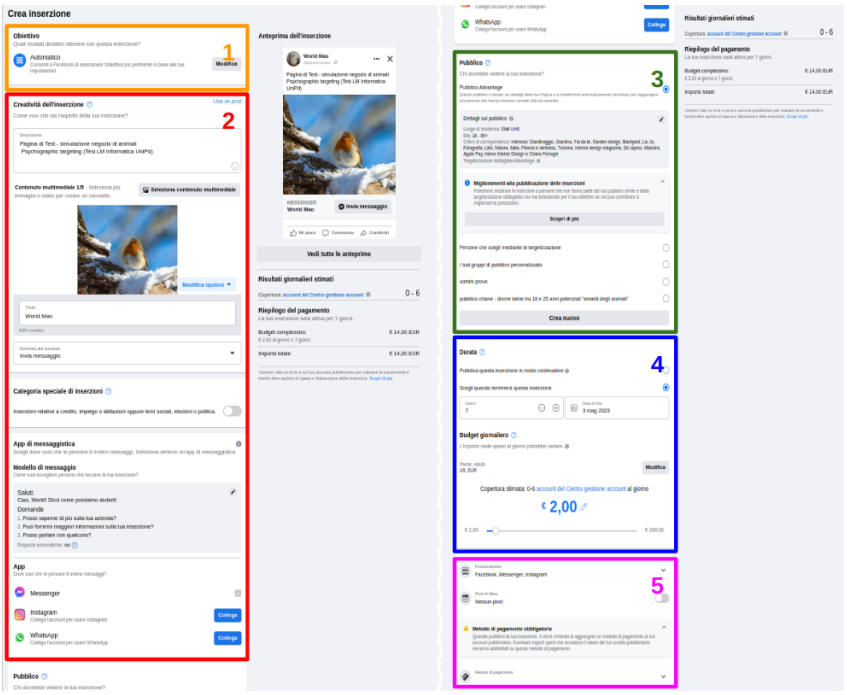
Alcuni vantaggi legati all'utilizzo della pubblicità mirata online sono il raggiungimento nell'immediato del pubblico corretto e la personalizzazione del target, un costo inferiore rispetto alla pubblicità tradizionale, risultati immediati con clienti "qui e ora", un ritorno dell'investimento maggiore grazie ai tassi di conversione più elevati e il *re-engagement* dei visitatori.
Va sottolineato che l'utilizzo dei dati personali a fini promozionali ha fatto sorgere sull'opinione pubblica, soprattutto negli ultimi anni, questioni etiche e morali [47] che tuttavia almeno per il momento sono poste in secondo piano rispetto ai vantaggi che la pubblicità mirata ha dimostrato come canale di marketing, anche grazie ai numerosi articoli e tutorial sul tema disponibili online con poco sforzo. Sono però state introdotte limitazioni, ad esempio per le campagne politiche.

5.1.2.1 Social networks

A seguire è presentato come viene gestita la pubblicità mirata online sui social networks *Facebook*, *Instagram* e *YouTube*.

***Facebook* e *Instagram*.** *Facebook* e *Instagram* sono piattaforme social, gestite e di proprietà dall'azienda *Meta* [102], con piattaforma pubblicitaria *Meta Business Suite* [99]. Per la creazione di una campagna pubblicitaria, Figura 5.1, un inserzionista deve crearsi una pagina aziendale su *Facebook* e successivamente, connettendosi a *Meta Business Suite*, compiere i seguenti passaggi (Figura 5.1a):

1. Scegliere l'obiettivo da promuovere;
2. Creare l'annuncio (*ad*);
3. Individuare il pubblico;
4. Definire la durata e il budget della campagna;
5. Scegliere dove attivare la campagna (*Facebook*, *Instagram* o entrambi);
6. Tracciare le proprie prestazioni con modifiche alla campagna quando necessario (Figura 5.1b con definizione della campagna, analisi dei costi e ricavi o valutazione dell'impatto sul pubblico).



(a) *creazione della campagna*

(b) *gestione della campagna*

Figura 5.1: *Come per un inserzionista è possibile creare e gestire una propria campagna pubblicitaria - Meta Business Suite.*

Uno dei passi fondamentali quando si desidera istanziare una campagna pubblicitaria, che può rendere vano qualsiasi importo di budget, è la scelta del pubblico potenziale. *Facebook* e *Instagram* mostrano automaticamente gli *ads* agli utenti che hanno maggiore probabilità di trovarli pertinenti. Inoltre l'inserzionista può indirizzare ulteriormente la pubblicizzazione con la selezione del pubblico (principale, personalizzato e simile) e sulla base di questo personalizzare il formato degli *ads*, Tabella 5.1.

Principale: *definizione in base a criteri quali posizione (città, comunità, paese d'affari); demografia (età, sesso, istruzione, lavoro); interessi (interessi e hobby); comportamento (acquisti precedenti e utilizzo del dispositivo); connessioni (inclusione/esclusione di utenti collegati alla pagina Facebook)*	
Immagine	Presentare prodotto e/o brand (file consigliati *jpg* e *png*)
Video	Catturare l'attenzione e spingere a eseguire un'azione in modo chiaro e semplice (durata max 15 sec., con audio consigliato)
Personalizzato: *definizione in base agli utenti che hanno già interagito con l'attività online o offline; liste di contatti, visitatori del sito e/o utenti dell'applicazione*	
Immagine	Presentare prodotto e/o brand (file consigliati *jpg* e *png*)
Video	Catturare l'attenzione e spingere a eseguire un'azione in modo chiaro e semplice (durata max 15 sec., con audio consigliato)
Carosello	Presentare prodotti che rimandano a diverse pagine di destinazione per mettere in evidenza caratteristiche del prodotto, creare storie coinvolgenti, spiegare un processo, evidenziare vantaggi (10 immagini/video con link a contenuto in un unico *ad*)
Raccolta	Esperienza interattiva con immagine di copertina, sotto la quale sono visualizzati più prodotti; per mostrare il proprio catalogo, trasformare le domande in vendite, esperienza di navigazione efficace su dispositivi mobile
Simile: *definizione in base a un pubblico di origine. Permette di raggiungere nuove persone i cui interessi sono simili a quelli del pubblico d'origine*	
Immagine	Presentare prodotto e/o brand (file consigliati *jpg* e *png*)
Video	Catturare l'attenzione e spingere a eseguire un'azione in modo chiaro e semplice (durata max 15 sec., con audio consigliato)

Tabella 5.1: *Personalizzazione degli ads in base al pubblico potenziale su Facebook e Instagram - Fonte: Inserzioni di Meta [101].*
Le interazioni online degli utenti vengono immagazzinate dai *Meta Pixel*, che registrano quando qualcuno esegue un'azione sul sito Web.

YouTube. Su *YouTube* è possibile realizzare pubblicità mirata online (configurazione con la piattaforma pubblicitaria *Google Ads* [54]), con la creazione di campagne sponsorizzate e più in generale aumentando la visibilità dei propri contenuti.
Per procedere a pubblicare un *ad* un inserzionista deve essere in possesso di un account *Google Ads*; quando soddisfatto tale requisito deve procedere con la creazione di un canale ed eseguire i successivi passi:

1. Caricare il video da usare per il proprio *ad*;

2. Avviare la campagna pubblicitaria;

3. Definire il pubblico a cui mostrare il contenuto;
4. Determinare il budget;
5. Personalizzare il formato degli *ads*;
6. Tracciare le proprie prestazioni con modifiche alla campagna quando necessario.

YouTube, gestita e di proprietà di *Google*, si contraddistingue per la sua piattaforma *user-friendly* che fa si che chiunque possa creare la propria campagna pubblicitaria, senza ricorrere a professionisti. Difatti *YouTube* include gli strumenti di *editing* che permettono anche all'inserzionista con nessun livello di esperienza in *videomaker* il montaggio di video online da browser.
La semplicità della piattaforma *YouTube* consente non solo di fare *advertising* ma anche di lavorare come *creator* e legalmente, quando e se occorre, comprare le visualizzazioni [42].
Figura 5.2 mostra come avviene in *Google Ads* la scelta di pubblicare un video pubblicitario su *YouTube*.

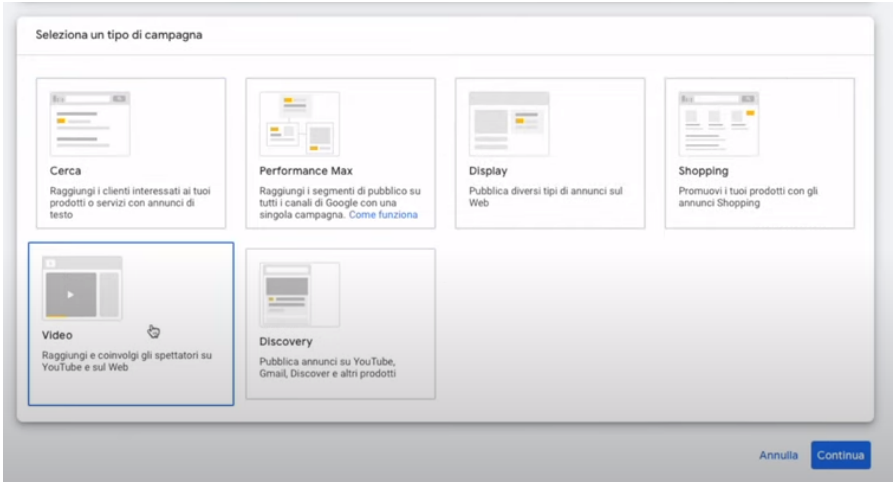


Figura 5.2: *Pubblicazione di un video ad su YouTube - Fonte: Digital Marketing Consulting [31].*

Un inserzionista può personalizzare il formato dei propri *ads*, Tabella 5.2, sulla base del proprio pubblico potenziale (per notorietà, per considerazione o per azione).

Notorietà: *cattura l'attenzione degli utenti e con l'ad mantiene l'attività al top*	
In-stream ignorabili	Apparizione in base al budget, a comparsa prima, durante e dopo la riproduzione di un video, ignorabile dall'utente
Bumper	Non ignorabile, attrae l'attenzione di un pubblico vasto con un messaggio breve, incisivo e memorabile (durata max 6 sec.)
In-stream non ignorabili	Non ignorabile, l'utente deve vedere per forza tutta la storia, a comparsa prima, durante o dopo la riproduzione di un video (durata max 15 sec.)
Masthead	Visualizzazione sulla parte superiore del feed, con il fine di raggiungere un gran numero di utenti in breve tempo. Da utilizzare per far conoscere un nuovo prodotto, servizio o raggiungere un ampio pubblico
Considerazione: *fa in modo che gli utenti per acquistare pensino subito all'azienda*	
In-stream ignorabili	Apparizione in base al budget, a comparsa prima, durante o dopo la riproduzione di un video, ignorabile dall'utente
Azione: *fa in modo che per gli utenti sia facile acquistare, iscriversi e altre azioni*	
Campagne video per azione	*Ads* su dispositivi mobile, desktop e TV, facilitando la generazione di più conversioni. Le campagne utilizzano formati in-stream ignorabili e in-feed video (durata max 10 sec. per video)
Annunci discovery	Impiego di immagini per raggiungere le persone mentre sfogliano i loro feed *YouTube*/*Gmail*.

Tabella 5.2: *Personalizzazione degli ads in base al pubblico potenziale su YouTube - Fonte: YouTube Advertising [163].*

Come già presentato per *Facebook* e *Instagram* anche su *YouTube* l'inserzionista può controllare se sta raggiungendo i suoi obiettivi; grazie a metriche che *Google Ads* mette a disposizione e che hanno come scopo il monitoraggio in tempo reale del pubblico potenziale [8] (per copertura del pubblico [9, 7], metriche di brand [6], metriche di performance [128, 116], visualizzazione ads [63] e misura d'interazione [5]).
Per ulteriori specifiche in merito alla scelta dell'obiettivo e sulle modalità di utilizzo del budget fare riferimento alla sezione §5.1.2.2.

5.1.2.2 Motori di ricerca

A seguire è presentato come viene gestita la pubblicità mirata online sul motore di ricerca *Google*.

Google. Il motore di ricerca *Google* offre servizi su:

* applicazioni e dispositivi (Ricerca e *YouTube* descritto in sezione §5.1.2.1);
* piattaforme (browser *Chrome* e sistema operativo *Android*);
* prodotti integrati in applicazioni e siti di terze parti (annunci pubblicitari, dati e analisi, e *Google Maps* incorporati).

Un annuncio pubblicitario *Google* può venire pubblicato su qualsiasi applicazione, piattaforma, servizio, partner di ricerca del motore; sulla base del budget e del *targeting* prescelto dall'inserzionista. Il modo con cui *Google* ne permette la creazione è mediante *Google Ads*, strumento che può utilizzare chiunque abbia intenzione di investire nella

pubblicità mirata online.
Google Ads per individuare il pubblico potenziale sfrutta le ricerche dei propri utenti, in questo modo è in grado di pubblicare un *ad* nel momento stesso in cui viene effettuata una ricerca che lo coinvolge. Inoltre l'inserzionista, durante la creazione dell'*ad*, definisce quale è l'obiettivo che vuole ottenere dalla pubblicazione, Figura 5.3, realizzabile con una delle tre macro-azioni di Tabella 5.3.

obiettivo	**azione**
Aumentarne le visite, le vendite online, le prenotazioni e le iscrizioni alla *mailing list*	Indirizzamenti al sito Web dell'inserzionista
Aumentare le chiamate dei clienti	Pulsante *click-to-call*
Facilitare l'individuazione della sede fisica del negozio	Incorporamenti di *Google Maps*

Tabella 5.3: *Azioni per realizzare gli obiettivi degli ads Google - Fonte: Google Ads [54].*

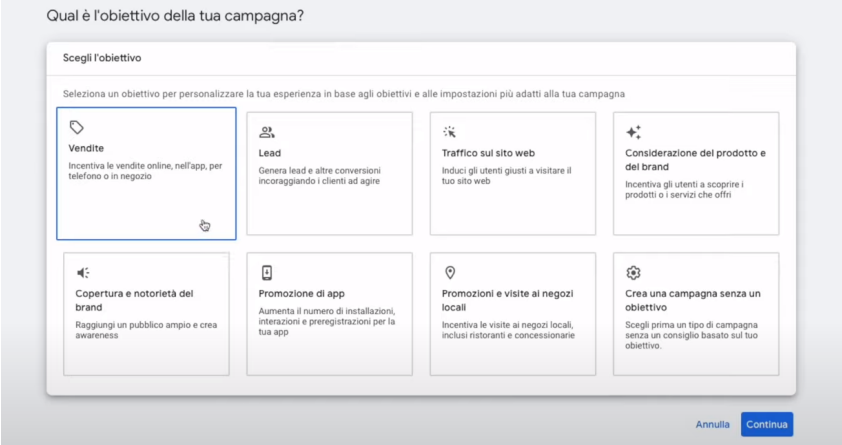


Figura 5.3: *Obiettivi degli ads Google - Fonte: Digital Marketing Consulting [31].*

È inoltre possibile per l'inserzionista scegliere il raggio (località, Stato o paese) di copertura del proprio *ad*, personalizzarne il formato Tabella 5.2, definire un budget anche sulla base di attività simili. Inoltre un costo viene conteggiato esclusivamente quando un utente fa click sull'annuncio.

5.1.3 Protezione degli utenti

Il social network che nel gennaio 2023 ha contato più accessi di utenti al mese è stato *Facebook* con 2'958 utenti attivi (seguito da *YouTube* con 2'514 utenti) [137]. Per quanto riguarda *Google* è stato stimato che controlla oltre 93% [135] del mercato globale dei motori di ricerca, elaborando 40'000 query di ricerche al secondo [64]. Conseguentemente come avviene la protezione dati degli utenti, enorme potenziale di mercato, è materia altamente sensibile oggetto di continue rivisitazioni da parte delle piattaforme media coinvolte.

5.1.3.1 Social networks

Facebook, *YouTube* e *Instagram* non condividono, se non con i propri gruppi di ricerca e team di analisti autorizzati, in alcun modo i *dati personali*[g] dei propri utenti; incluse le operazioni che questi compiono all'interno dei social networks a meno di specifiche concessioni. I dati quando trasmessi da social a terzi, lo sono solo dopo essere stati sottoposti a metriche di aggregazione o a procedura di *hashing* (quest'ultimo da *Meta* durante la creazione di un pubblico personalizzato da una elenco di clienti). Quando, invece, devono essere trasmessi a enti esterni a scopo di ricerca per l'innovazione e il bene della società:

* *Meta*, fornisce previa analisi e autorizzazione della domanda, l'accesso in chiaro a tutti i suoi dati raccolti (da fonti pubbliche, gruppi di professionisti e no-profit) [100]; tutelando però sempre la privacy dei propri utenti che non potranno mai essere riconosciuti.

* *YouTube*, da la possibilità ai ricercatori di partecipare a un proprio programma interno [164], con cui è possibile avere accesso ridimensionato alla raccolta di dati.

In aggiunta vengono rilasciati al pubblico, periodicamente dai social networks, datasets a fine di ricerca come:

* I datasets rilasciati da *Data for Good* [98] di *Meta*, con licenza *free to use* che in prevalenza riguardano la densità della popolazione in diverse aree del mondo [97], e che possono venire utilizzati per analisi sulla popolazione e sulla densità geografica.

* *YouTube-8M* [3], set di dati video rilasciato da *YouTube* in collaborazione con *Google*. Con lo scopo di accelerare la ricerca sulla comprensione dei video su larga scala, l'apprendimento della rappresentazione, la modellazione dei dati rumorosi, l'apprendimento del trasferimento e gli approcci di adattamento del dominio.

I dati pubblici degli utenti sono invece liberamente accessibili e non subiscono alcuna limitazione; tuttavia, dalle nostre ricerche è emerso come sono molto difficili da estrarre dalle piattaforme *Meta* e *Google Ads*, anche con l'utilizzo delle *API*[g] apposite (per i lunghi tempi di latenza), e in ogni caso si limitano a prelevare le informazioni pubbliche solo degli utenti che aderiscono al canale o alla pagina *Facebook*.

5.1.3.2 Motori di ricerca

Google raccoglie informazioni anche su utenti che non hanno un account. Questo significa che il motore è in possesso di dati personali di una buona fetta della popolazione mondiale (2 miliardi di utenti solo stimando *Android*); un potenziale enorme se considerati gli studi di mercato, le analisi pubblicitarie e le profilazioni dei clienti possibili. La politica di *Google* [55] per gestire la questione, evitando polemiche e scandali, è simile a quelle messe in atto dai social networks; per cui non vi è alcuna condivisione delle informazioni personali degli utenti con aziende, organizzazioni e individui esterni a *Google*.
A tale regola sono, tuttavia, previste delle eccezioni che coinvolgono: gli amministratori di dominio che possono accedere alle informazioni memorizzate negli accounts dei loro utenti (come, ad esempio, *Gmail*); l'elaborazione esterna da parte di affiliate di *Google* (come, ad esempio, fornitori di servizi per gestire i *data center*); per motivi legali (come,

ad esempio, per soddisfacibilità di legge, regolamenti e/o procedura legale).
Per quando riguarda la ricerca accademica, anche *Google* mette a disposizione dei propri datasets anonimi; inoltre permette sempre la collaborazione tra Università e i suoi gruppi di ricerca, istanziando quando necessario borse di studio o piattaforme per l'elaborazione dei dati (*Google Cloud*) [53], [52].
In ogni caso è sempre possibile per il motore di ricerca condividere le informazioni pubbliche con i suoi partner, inclusi gli inserzionisti.

5.1.4 Conclusioni

Dalle nostre analisi su funzionamento e protezione degli utenti, emerge come la pubblicità mirata online sia uno strumento su cui i Media digitali stanno investendo molto per attrarre sempre più aziende e possibili partner. Complice di tale meccanismo è sicuramente il grande bacino di utenti che possono venire raggiunti facilmente e comodamente solo connettendosi online, catalizzante sia per le aziende che cercano di vendere i loro prodotti che per i potenziali clienti che desiderano un acquisto.
In tale contesto assume vitale importanza come vengono gestiti i dati privati degli utenti coinvolti, che non possono essere divulgati liberamente (*Regolamento europeo GDPR* [30]) in modo da tutelare la collettività e il singolo da possibili illeciti. A oggi su tale fronte, sia *Meta* che *Google Ads*, non divulgano in alcun modo direttamente al loro esterno i dati personali dei propri utenti (solo mediante metriche aggregate, procedure di *hashing* e dati anonimati); questo probabilmente sull'onda anche degli scandali come *Cambridge Analytica*.

Per l'applicazione del *targeting* psicografico, da come appare nell'analisi, nessuno dei Media digitali analizzati include alcuna funzionalità che può portare alla sua applicazione diretta in quanto è mancante la fase di profilazione psicologica del pubblico potenziale (si vedano sezioni §5.1.2.1 e §5.1.2.2). Inoltre la protezione dei dati degli utenti messa in atto dalle piattaforme pubblicitarie rende impossibile un'integrazione del *targeting* psicografico direttamente profilando gli utenti dei media (si vedano sezioni §5.1.3.1 e §5.1.3.2).
Tuttavia individuiamo in *Facebook* la possibilià di realizzare il *targeting* psicografico in maniera indiretta, descritto nelle sezioni §5.2 e §5.3, utilizzando la creazione di un pubblico personalizzato per esempio da una lista di contatti (Tabella 5.1, i quali tratti possono essere precedentemente dedotti dall'inserzionista profilando i propri clienti con tecniche di *Natural Language Processing*) per la profilazione psicologica e la creazione di *ads* specifici per il pubblico personalizzato per gli interventi psicologicamente informati (si veda sezione §5.1.2.1).
Questo consentirebbe grazie alla combinazione tra Marketing, Media digitali e Intelligenza Artificiale la realizzazione di uno strumento, liberamente accessibile e capace di coinvolgere un ampio pubblico senza violazioni della privacy, che si presta a essere utilizzato non solo a scopi monetari ma anche per risolvere questioni etiche-sociali, come il *gender gap* in ambito occupazionale descritto in sezione §5.3.2.1.

5.2 Studio di fattibilità del *targeting* psicografico

La seconda parte della nostra analisi sulla pubblicità mirata online ha riguardato uno studio di fattibilità del *targeting* psicografico per comprendere come è possibile applicarlo all'interno dei Media digitali.

5.2.1 Metodologia

Ci siamo occupati di esaminare alcune possibili applicazioni del *targeting* psicografico all'interno dei media; con focus particolare la sua fattibilità in *Facebook*.
Abbiamo scelto di concentrarci su *Facebook* in quanto a differenza di *Google Ads* permette l'individuazione, con *Meta Business Suite*, di un pubblico potenziale (con pubblico personalizzato) sulla base di un pubblico fornito dall'inserzionista stesso. Tale pubblico personalizzato può dunque essere basato su specifici tratti di personalità e creare così la fase di profilazione psicologica fondamentale per la realizzazione del *targeting* psicografico.

5.2.2 *Targeting* psicografico su *Gmail*

Come spiegato in sezione §5.1.3.2 è possibile, da politica di *Google*, per un amministratore di dominio accedere alle informazioni contenute all'interno degli accounts dei propri utenti. Abbiamo dunque esaminato tale atto quando il servizio di *Google* è *Gmail*.
In un'azienda o un ente pubblico che si avvale dei servizi *Google* per creare il proprio dominio di posta elettronica; gli amministratori del dominio possono in qualsiasi momento avere accesso al contenuto dei messaggi scambiativi al suo interno e a qualsiasi altra informazione collegata agli accounts. Ciò implica non solo un controllo diretto, per esempio, di dirigente verso dipendenti, ma anche la possibilità con le attuali tecniche di *Natural Language Processing* di estrarne il contenuto e, per esempio, profilare i possessori degli accounts psicologicamente; come mostrato in questa tesi in Capitolo §4. Una conseguenza a ciò, che abbiamo ritenuto fattibile, è l'attualizzazione da parte degli amministratori o di chi gli ha delegati (dirigente) di interventi psicologicamente informati, a beneficio (o danno) degli utenti (dipendenti e clienti).
Quanto appena definito sopra è legale solo con apposite autorizzazioni dei soggetti coinvolti (*art.4 Regolamento europeo GDPR* [30]); tuttavia tale discussione prova anche che è possibile compiere questa azione "all'oscuro" sfruttando esclusivamente il fatto di essere amministratori di un dominio *Google*.

5.2.3 *Targeting* psicografico su *Facebook*: caso studio *World Mao*

Per valutare l'applicabilità del *targeting* psicografico su *Facebook* dovevamo prima di tutto procedere con la creazione di una pagina aziendale; *World Mao*[1] Figura 5.4.

[1] https://www.facebook.com/animal.world.mao.

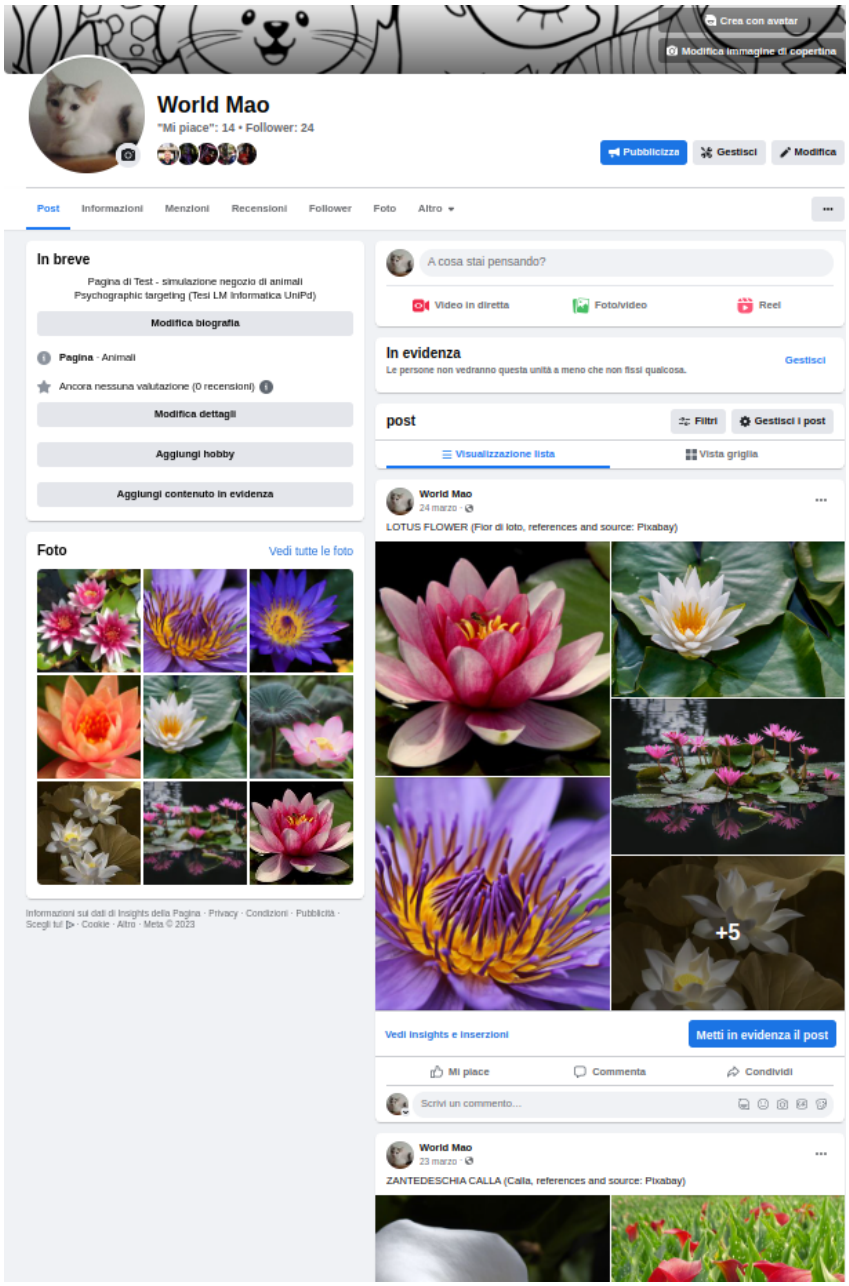


Figura 5.4: *Pagina di post - pagina World Mao.*

Una volta creata la pagina abbiamo:

1. **Definito lo scopo della pagina Web:** essendo una pagina di *Test*, con nessun fine reale di pubblicità mirata, abbiamo scelto di simulare un negozio di animali;

2. **Individuato i nostri *followers*:** abitualmente il pubblico viene individuato, come spiegato in sezione §5.1, mediante tecniche pubblicitarie; essendo il nostro un lavoro di studio non finanziato e a durata limitata abbiamo creato il nostro

pubblico tra gli studenti di Informatica, Data Science e Psicologia dell'Università degli Studi di Padova disponibili a seguire la pagina e in possesso di un account *Facebook*.

3. **Proceduto con la simulazione:** la simulazione si è svolta in due fasi e ci ha permesso di comprendere il funzionamento di una pagina *Facebook* pubblicitaria. Durante la prima fase, dal 12/01/2023 al 24/03/2023, abbiamo proceduto con la pubblicazione giornaliera in *World Mao* di almeno un post contenente immagini di animali o piante; durante la seconda fase ci siamo occupati di svolgere dei test su alcune funzionalità di pubblicità mirata offerte della piattaforma.

4. **Valutato la fattibilità del *targeting* psicografico**: abbiamo analizzato come fosse possibile con le attuali tecniche di pubblicità mirata, integrate all'interno della piattaforma pubblicitaria *Meta*, tracciare il proprio pubblico potenziale anche per tratti di personalità *Big Five*.

5.2.3.1 Prima fase: risultati ottenuti dalle pubblicazioni

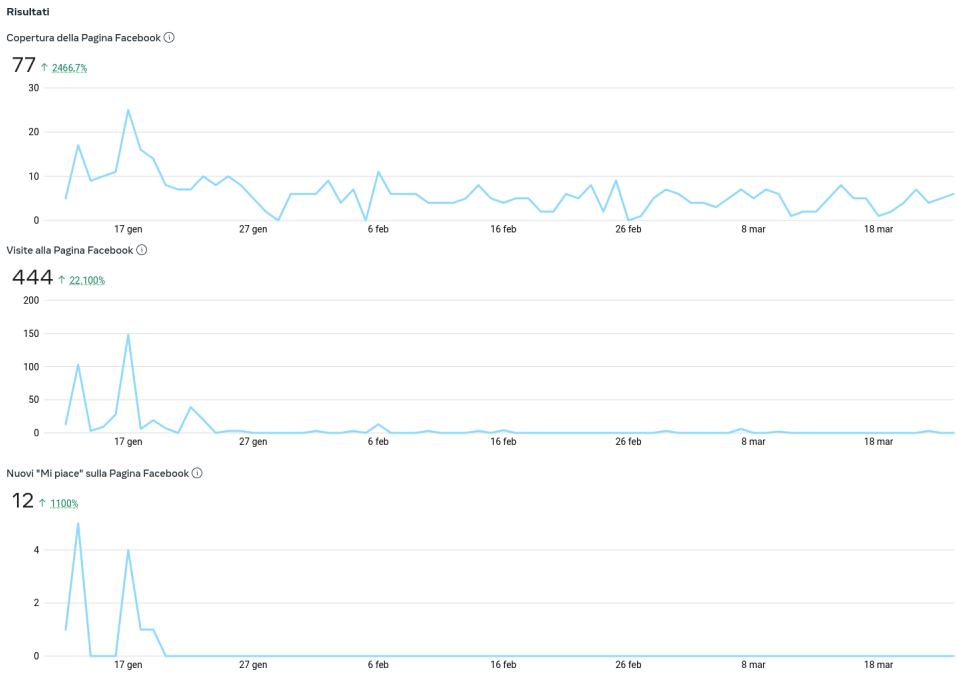


Figura 5.5: *Copertura, Visite alla Pagina e "Mi Piace" dal 12/01/2023 al 24/03/2023 - pagina World Mao.*

Dalle pubblicazioni compiute durante la prima fase della simulazione abbiamo ottenuto i risultati di Figura 5.5. Da questi è evidente come la pagina abbia collezionato la maggior parte di visite, dei "Mi Piace" e la copertura (numero di utenti che hanno visualizzato i contenuti all'interno della pagina) in prevalenza nella prima parte del periodo.
La situazione è giustificata dalla difficoltà di far visualizzare contenuti di una pagina

Facebook senza apposite strategie di pubblicità mirata; ecco che di conseguenza le nostre interazioni sono risultate scarse e concentrate soprattutto quando *World Mao* ha collezionato la maggioranza del suo pubblico attuale (12/01/2023 - 22/01/2023, 24 *followers* e 14 "Mi Piace").
Tali risultati possono essere considerati una prova della necessità e validità delle strategie offerte da *Meta Business Suite* e *Google Ads* nel campo della pubblicità mirata online senza delle quali il giro d'affari mondiale di 1'255 miliardi del triennio 2020-2022 [92] non sarebbero mai stati raggiunti.

5.2.3.2 Seconda fase: selezione del pubblico potenziale

Nella seconda fase della simulazione ci siamo concentrati non più sulla pubblicazione di annunci pubblicitari, perchè al di fuori degli obiettivi della nostra ricerca ma nel comprendere come avviene la targetizzazione.
Come presentato in sezione §5.1.2.1 in *Facebook* per attrarre pubblico alla propria pagina Web, si può optare per una targetizzazione per pubblico principale (o salvato), personalizzato, simile, o anche una combinazione tra più. Figura 5.6 mostra come si presenta la scelta del pubblico da *Meta Business Suite* sulla pagina *World Mao*.

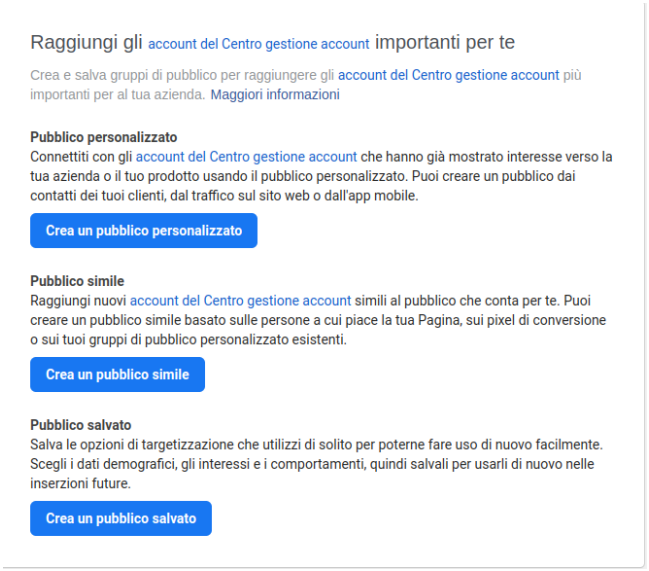


Figura 5.6: *Targetizzazione sulla base del pubblico - pagina World Mao.*

I risultati che abbiamo ottenuto durante i nostri test, indicano che è possibile per un inserzionista creare il proprio pubblico sulla base di queste targetizzazioni, senza però avere alcun dato individuale degli utenti se non in forma aggregata. A prova di ciò in Figura 5.7 mostriamo come abbiamo definito un pubblico principale composto da donne "amanti degli animali" con lingua di origine latina; Figura 5.8 mostra tutti i pubblici potenziali che abbiamo generato durante questa fase e come sia impossibile per un inserzionista ottenere informazioni personali sul proprio pubblico.

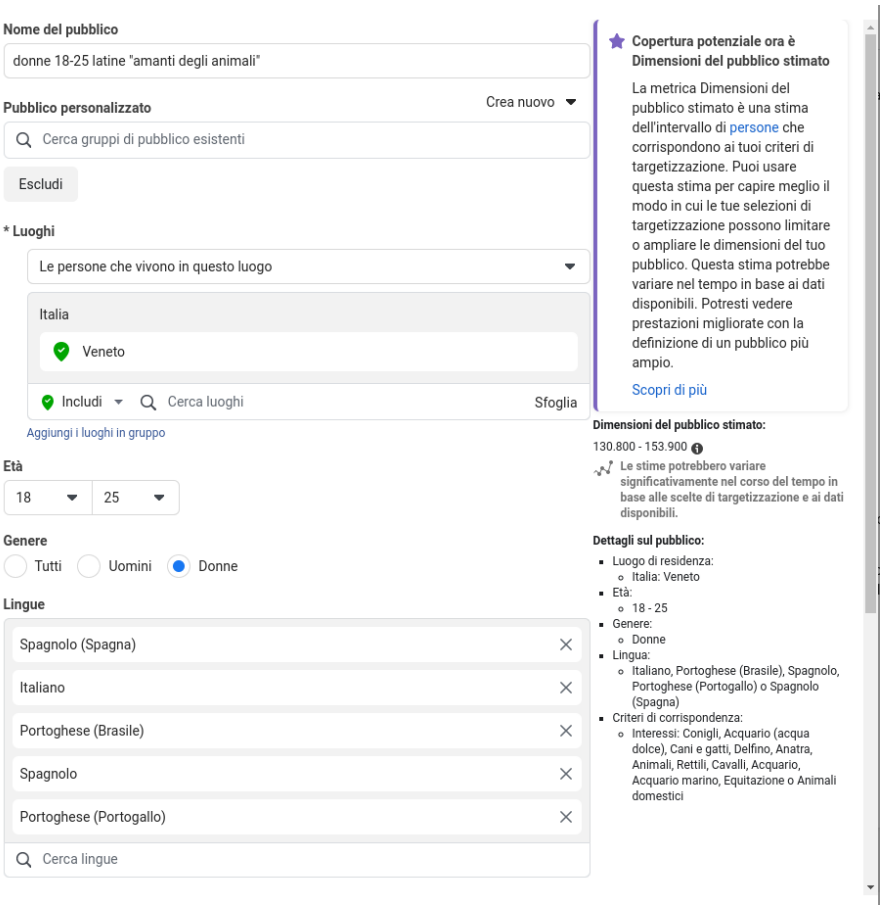
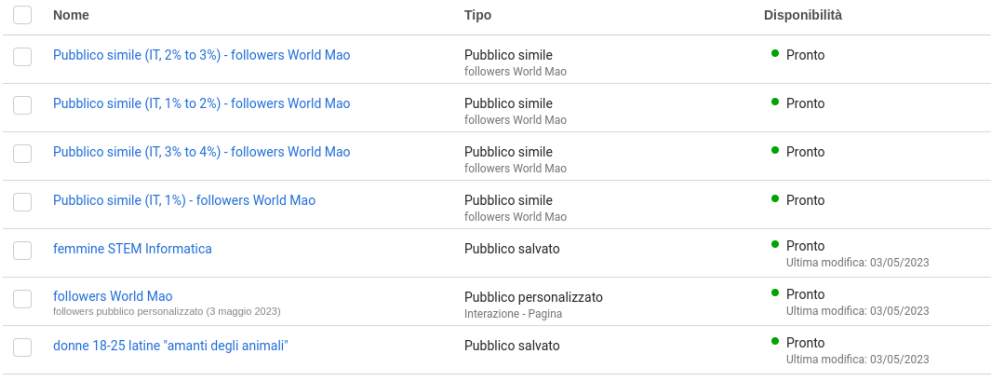


Figura 5.7: *Pubblico principale di donne tra i 18-25 anni di lingua latina (italiana, spagnola o portoghese) residenti in Veneto (Italia) - pagina World Mao.*
Interessi: Conigli, Acquario (acqua dolce), Cani e gatti, Delfino, Anatra, Animali, Rettili, Cavalli, Acquario, Acquario marino, Equitazione o Animali domestici. Utilizzando la voce *Pubblico personalizzato* è possibile anche targetizzare un pubblico personalizzato e/o simile.



Nome	Tipo	Disponibilità
Pubblico simile (IT, 2% to 3%) - followers World Mao	Pubblico simile followers World Mao	Pronto
Pubblico simile (IT, 1% to 2%) - followers World Mao	Pubblico simile followers World Mao	Pronto
Pubblico simile (IT, 3% to 4%) - followers World Mao	Pubblico simile followers World Mao	Pronto
Pubblico simile (IT, 1%) - followers World Mao	Pubblico simile followers World Mao	Pronto
femmine STEM Informatica	Pubblico salvato	Pronto Ultima modifica: 03/05/2023
followers World Mao followers pubblico personalizzato (3 maggio 2023)	Pubblico personalizzato Interazione - Pagina	Pronto Ultima modifica: 03/05/2023
donne 18-25 latine "amanti degli animali"	Pubblico salvato	Pronto Ultima modifica: 03/05/2023

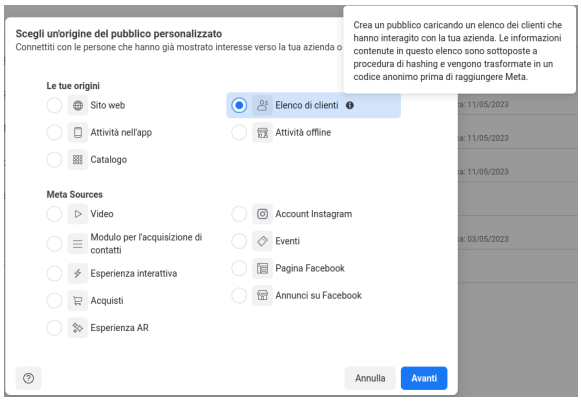
Figura 5.8: *Targetizzazioni generate durante la fase due - pagina World Mao.*
Il pubblico principale *femmine STEM Informatica* è a dimensione fissa, lo abbiamo creato dai *followers* della pagina selezionando donne tra i 18-35 anni con qualche interesse per l'Informatica (Interessi: Scienza, Computer, Giochi online, Ingegneria, Agenzia Spaziale Europea, UFO Files, Matematica, Web design, Java, Elettronica, Python, NASA, Programmazione informatica o C++, Campo di studio: Computational engineering o HTML, Titolo professionale: Data science); il pubblico personalizzato *followers World Mao* è incrementale sui *followers* della pagina; il pubblico principale *donne 18-25 latine "amanti degli animali"* è a dimensione fissa; i pubblici simili a 4 segmenti (più elevata è la percentuale e più il pubblico simile differisce dall'origine) sono a dimensione incrementale e li abbiamo individuati da *followers World Mao*.

5.2.3.3 Fattibilità: estensione con *targeting* psicografico

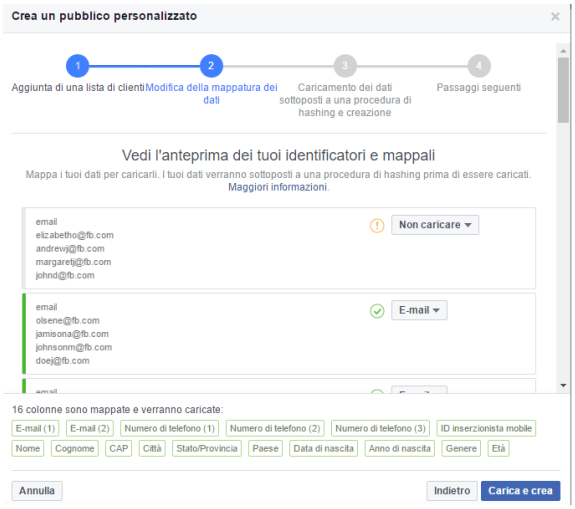
Il *targeting* psicografico non è un'opzione di pubblicità mirata online attualmente integrata all'interno delle piattaforma dei Media digitali. Questo vale anche per *Meta Business Suite* (si veda sezione §5.1); la cui funzione più simile alla psicografia consiste nella possibilità di selezionare un pubblico potenziale su pubblico principale per comportamenti (come, per esempio, abitudini d'acquisto, anniversari, attività digitali e conoscenze).
Per studiare un'integrazione del *targeting* psicografico all'interno della piattaforma abbiamo testato due direzioni: elenco di clienti e *followers* della pagina.

Elenco di clienti. Abbiamo assunto, nella nostra simulazione di negozio di animali, di essere in possesso di un elenco di clienti. Essendo nostri clienti, è ragionevole pensare di avere già avuto qualche interazione con loro e per questo per ognuno conoscere almeno nome, cognome ed e-mail. Detenendo gli indirizzi di posta elettronica abbiamo ipotizzato di avere avuto con loro uno scambio di corrispondenza in linguaggio naturale e, come già dimostrato, dal linguaggio naturale si è in grado di ricavare i tratti di personalità *Big Five* di chi ne fa uso. È inoltre possibile, a partire dall'indirizzo e-mail, risalire ai profili associati sui social networks (per esempio su *Twitter* si può risalire se l'email è pubblica all'utente proprietario inserendola nell'apposita barra di ricerca del media), e da quelli estrarre le informazioni pubbliche inclusi alcuni post in linguaggio naturale utili alla predizione del tratto. Conseguentemente abbiamo potuto assumere di essere capaci, con adeguate autorizzazioni da parte dei nostri clienti, di generare delle liste ad hoc composte dalle e-mail dei clienti che rientravano in un particolare

profilo di personalità o per combinazione di tratti selezionati nel pubblico che volevamo raggiungere.
È possibile fornire ognuna di queste liste clienti, Figura 5.9 a *Meta Business Suite* (Figura 5.9a) la quale mediante procedura di *hashing*, fa si che le informazioni personali dei clienti nella lista (Figura 5.9b) non siano visibili alla piattaforma pubblicitaria, restituendo un pubblico personalizzato composto dagli accounts *Facebook* degli utenti a cui appartengono le e-mail della lista clienti con i tatti di personalità selezionati (Figura 5.9c).



(a) *selezione modalità "Elenco di clienti"*



(b) *selezione e-mail da convertire in utenti Facebook - Fonte: Come caricare la tua lista clienti su Facebook [143]*

(c) *pubblico personalizzato generato da una lista di e-mail - Fonte: Come caricare la tua lista clienti su Facebook [143]*

Figura 5.9: *Generazione pubblico personalizzato da un elenco di clienti.*

Il processo di quanto appena descritto, rappresentato graficamente in Figura 5.10, permette a qualsiasi inserzionista in possesso di liste clienti per tratti di personalità di indirizzarvi delle pubblicità mirate (a immagine, video, carosello o raccolta, Tabella 5.1) specifiche per i tratti. Questo rappresenta la realizzazione, mediante *Facebook*, del *targeting* psicografico.
Inoltre sul pubblico personalizzato ottenuto è possibile applicare un'ulteriore selezione mediante targetizzazione dettagliata (pubblico principale) o generare un pubblico simile, o combinare i due approcci.

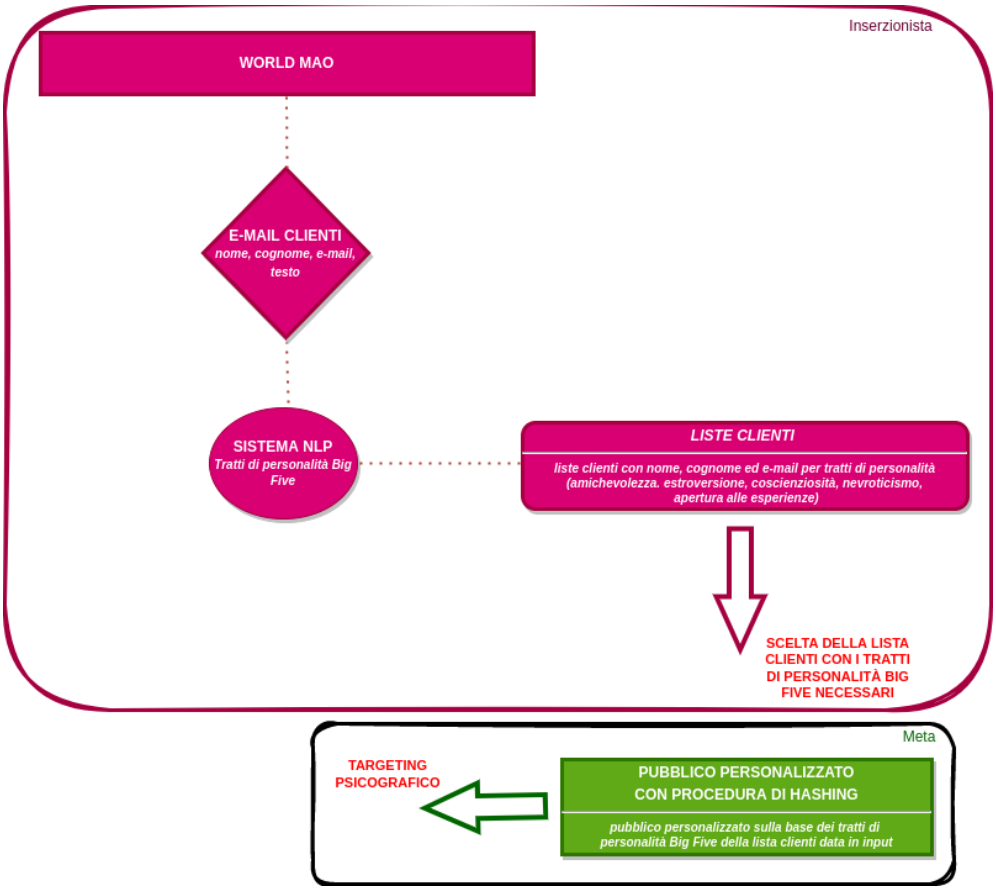


Figura 5.10: *Applicazione del targeting psicografico - pagina World Mao.*

Followers. Un'altra strategia che abbiamo cercato di testare è stata quella di capire se sia possibile un'integrazione del *targeting* psicografico utilizzando i *followers* della pagina di un inserzionista. Tuttavia questo non ci è stato possibile in alcun modo, nemmeno cercando di individuare i *followers* senza targetizzazione, come invece è avvenuto per i pubblici in Figura 5.8; a conferma effettiva che le informazioni e le azioni degli utenti sono personali, se non sono questi ultimi a decidere altrimenti.

5.2.4 Conclusioni

Il nostro studio di fattibilità sul *targeting* psicografico ha dato esito positivo, emerge dunque come è possibile integrare all'interno di una piattaforma pubblicitaria online delle funzionalità indirette, di influenza della scelta sulla base del target, per cui non è stata progettata. Questo dovrebbe spingere le Aziende, gli Enti pubblici e i Professionisti a non limitarsi a utilizzare esclusivamente gli strumenti offerti dalle piattaforme, lasciando ai Media digitali l'amministrazione del mercato online; ma farsi carico loro stessi di comprendere il comportamento del proprio pubblico potenziale e di decidere di conseguenza la strategia di pubblicità mirata online più adeguata.

5.3 Applicazione nella società

Nell'ultima parte dell'analisi sulla pubblicità mirata online abbiamo esteso quanto appreso in merito al *targeting* psicografico, e descritto nelle sezioni §5.1 e §5.2, per risolvere i problemi legati al *gender gap* in ambito STEM.

5.3.1 Metodologia

Prima abbiamo proceduto a studiare il caso più generale del basso numero di presenze nelle occupazioni, per poi concretizzarlo nelle discipline STEM.

5.3.2 Come alleviare il basso numero di presenze in alcuni ambiti occupazionali (ad alta specializzazione)

Per cercare di attrarre la scelta lavorativa delle persone su specifici ambiti professionali (con una domanda di lavoratori superiore ai laureati o come nel caso delle donne in area STEM) ad alta specializzazione e che richiedono studi universitari abbiamo individuato nel *targeting* psicografico di studentesse e studenti una possibile strategia, volta a:

1. Influenzarne, rendendo chiaro il fine per non cadere in situazioni di manipolazione, la scelta universitaria se con tratti di personalità e interessi in linea con quelli dell'occupazione a cui la scelta universitaria viene associata;

2. Attrarne con tratti di personalità differenti ma interessi sufficientemente conformi alle personalità associate all'occupazione. Questo lo definiamo necessario per evitare ambienti con un'elevata polarità dei tratti [17, 81] e/o situazioni di iniquità; promuovendo invece un ambiente di studio e successivamente occupazionale diversificato, adattivo e di performance collettiva [12].

La nostra scelta di targetizzare studentesse e studenti, per limitare le disparità occupazionali ad alta specializzazione, risiede sulla visione delle occupazioni come conseguenza diretta dell'istruzione. Per cui per riuscire a modificare specifici comportamenti e scelte occupazionali riteniamo che sia necessario, in questo caso, agire sulle scelte delle studentesse e degli studenti in ambito pre-universitario (ultimi tre anni delle scuole superiori, fascia 16-18 anni) in modo da precedere la loro decisione di "cosa fare nella vita" e limitare l'influenza dagli stereotipi (in ambito occupazionale soprattutto di genere) che possono permeare un ambiente di studio e lavorativo ad altà intensità [66].

Il mezzo sul quale abbiamo teorizzato sia fattibile applicare il *targeting* psicografico sono i Media digitali quali *Facebook* e *Instagram* (potenzialmente estendibile anche

su altre piattaforme maggiormente in uso dai giovani adulti, come *TikTok* [136]) le piattaforme che dalle nostre analisi consentirebbero di giungere più facilmente ad attrarre potenziali studenti e futuri lavoratori di tipo (1) e (2), cogliendone i rispettivi vantaggi.
Inoltre l'uso dei tratti di personalità non risente, essendo caratteristiche soggettive, degli stereotipi della società. Questo quindi permetterebbe l'attrazione di un pubblico potenziale di nuovi lavoratori per disciplina privo di stereotipi e con una maggiore diversificazione di personalità, che incentiverebbe le discipline occupazionali a uscire da schemi e dogmi fissati.

5.3.2.1 *Gender gap* in occupazioni STEM

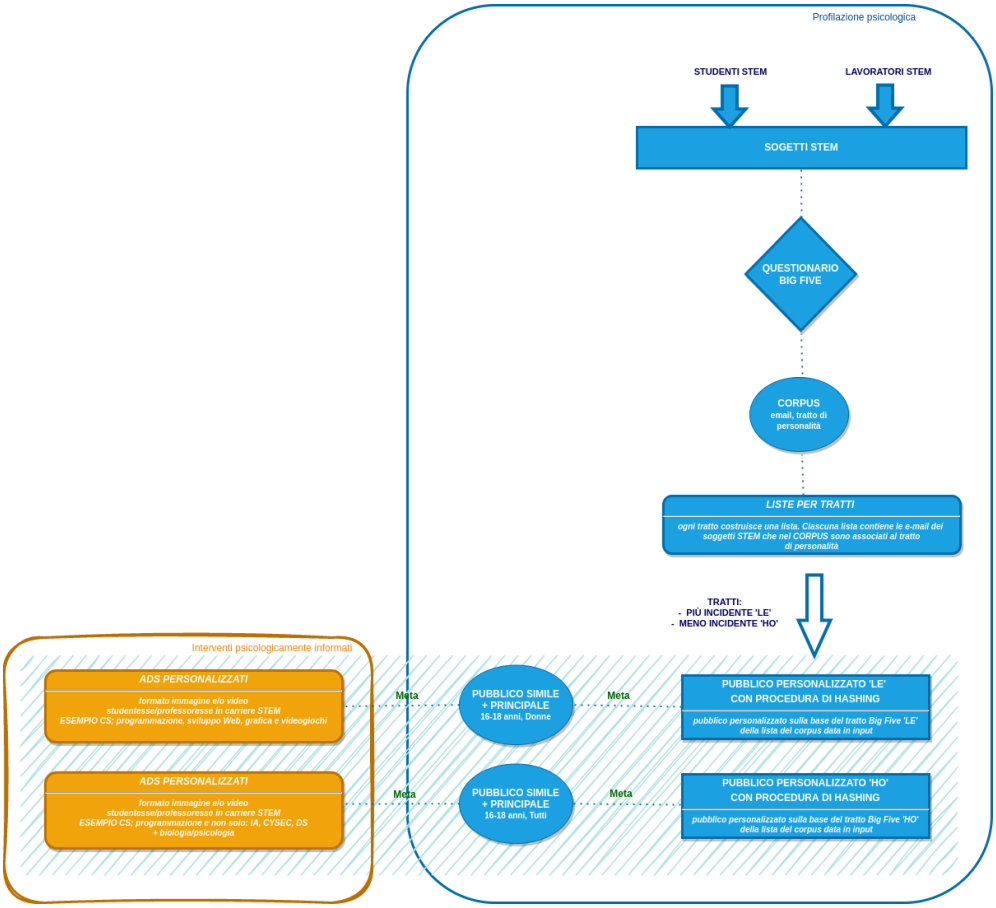


Figura 5.11: *Applicazione del targeting psicografico per alleviare il gender gap occupazionale STEM.*

In Figura 5.11 presentiamo graficamente la nostra strategia di *targeting* psicografico, comprensiva di interventi psicologicamente informati, per tentare di alleviare il *gender gap* occupazionale STEM.

Per cercare di alleviare i disequilibri presenti in ambito STEM ci siamo focalizzati

nell'attrazione di studentesse STEM introverse, e nella ricerca di potenziali studentesse e studenti STEM con tratti differenti da quello di maggiore correlazione (Tabella 4.22). Tale procedura potrebbe essere messa in atto dalle Università, dalle Aziende e dagli Enti pubblici con l'ausilio del *targeting* psicografico, limitando così il *gender gap* e la nascita di ambienti STEM poco performanti, nel modo seguente:

1. **Profilazione psicologica:** individuare un campione di soggetti STEM (studenti e lavoratori per evitare correlazioni implicite tra tratti di personalità e periodo di carriera) con le rispettive e-mail (oppure nome e cognome). I partecipanti vanno sottoposti a questionari *Big Five*, appositamente costruiti in grado di rilevare anche le *digital footprints* con il fine di ottenere un *corpus* delle impronte privo di *bias* digitali, di media e di modello. Il *corpus* va successivamente suddiviso per tratti di personalità.
   Successivamente fare uso della piattaforma pubblicitaria di *Facebook* (dal nostro studio di fattibilità):
   * Sulle e-mail dei soggetti STEM con il tratto di personalità di maggioranza del *corpus* applicarvi la tecnica dell'elenco di clienti, e ottenere in questo modo un pubblico personalizzato, da cui individuare un pubblico simile con targetizzazione dettagliata (pubblico principale) per parametri d'età 16-18 e genere femminile;
   * Sulle e-mail dei soggetti STEM con i tratti di personalità meno incidenti all'interno del *corpus* applicarvi la tecnica dell'elenco di clienti, e ottenere in questo modo un pubblico personalizzato, da cui individuare un pubblico simile con targetizzazione dettagliata (pubblico principale) per parametri d'età 16-18 e entrambi i generi.
2. **Interventi psicologicamente informati:** una volta individuato il pubblico potenziale, sulla base del tratto di personalità e genere del target, indirizzarvi dei messaggio pubblicitari sviluppati ad-hoc con il fine di attrarre allo studio delle materie STEM.

5.3.2.2 Messaggi pubblicitari sviluppati ad hoc

Come realizzare i messaggi pubblicitari, dell'intervento psicologicamente informato, è al di fuori delle competenze di questa tesi; tuttavia abbiamo ritenuto di doverne approfondire alcun aspetti. Di seguito ne illustriamo alcuni riferiti all'ambito STEM.

Pubblico potenziale. Il pubblico potenziale STEM è un pubblico principale definito su un pubblico personalizzato simile; dunque sconsigliamo l'utilizzo di *ad* come Raccolta e Carosello (Tabella 5.1) in quanto troppo impegnativi per un soggetto "che non ci conosce".

Congruenza con i tratti di personalità *Big Five*. Il pubblico potenziale di studentesse e studenti STEM prevediamo sia composto da soggetti introversi (tratto di personalità più incidente) e aperti alle esperienze (uno dei tratti di personalità meno incidenti).
Per la teoria di congruenza e di scaristà con i tratti *Big Five* [156] che definisce gli individui maggiormente attratti dai messaggi pubblicitari in linea con i propri tratti di personalità soprattutto quando questi ultimi non saturano un determinato ambiente, consigliamo per il tratto introverso di realizzare *ads* che evidenziano gli aspetti di

tranquillità, riservatezza e silenziosità legati a un'occupazione STEM (per esempio, un programmatore di fronte al suo computer); per il tratto invece di apertura consigliamo la realizzazione di *ads* che mettano in luce la dinamicità dell'ambiente (per esempio puntare a mostrare una molteplicità di figure e possibilità: programmatore ma anche analista dei dati, lavoro in azienda ma anche accademico e d'insegnamento).

Congruenza con il genere. Se l'*ad* è rivolto a un pubblico femminile, consigliamo di focalizzare il messaggio verso un lavoro socialmente utile (Tabella 4.22) e far veicolare il messaggio da figure femminili che assumono il ruolo di mentore [123, 32]. Invece se il messaggio è rivolto a maschi non tralasciare completamente il ruolo femminile, per incentivare la nascita del collegamento STEM-donna nella società; però prediligere evidenziare aspetti manuali e ingegneristici legati ai lavori STEM.

Capitolo 6

Conclusioni

In questo capitolo presentiamo le nostre conclusioni a questo lavoro di tesi e le sue possibili future applicazioni.

6.1 Considerazioni

Questo lavoro di tesi evidenzia che esistono delle relazioni tra genere, personalità e occupazione dimostrando che tratti indipendenti dal volere dell'individuo sono in grado di influenzarne le scelte. Nello specifico abbiamo individuato delle relazioni tra le occupazioni STEM e i tratti di personalità di introversione e nevroticismo, con forza maggiore nei maschi rispetto alle femmine. Abbiamo anche rilevato delle differenze tra le discipline STEM, con l'Informatica caratterizzata nuovamente da introversione e nevroticismo soprattutto maschile, e la Matematica che invece si dimostra composta da soggetti, di entrambi i generi, con stabilità emotiva e amichevolezza. Inoltre la presenza del nevroticismo in Informatica, ma non in Matematica, ci ha fatto ipotizzare che questi sia motivato più dalla programmazione che dal genere di appartenenza.
Tuttavia abbiamo associato i tratti rilevati anche alla piattaforma media sul quale abbiamo svolto l'analisi, in quanto l'utilizzo di social networks risulta da studi precedenti associato in prevalenza a soggetti tendenti all'introversione e al nevroticismo. Per questo motivo ci aspettiamo che nella realtà i nostri risultati siano mediati da questo fattore.
All'inizio del nostro lavoro ci eravamo chiesti se esistesse qualche relazione tra alto nevroticismo, le femmine e le carriere STEM e se fosse questo elemento, generando sentimenti di non appartenenza, che limitasse la partecipazione delle donne alle STEM. Tuttavia non abbiamo individuato alcuna relazione. Al contrario il nevroticismo che è emerso nelle femmine, considerando esclusivamente l'impatto della personalità sul genere, nelle occupazioni STEM è venuto a mancare indicando forse una qualche relazione tra le donne nevrotiche e lo scarso uso dei social networks, soprattutto se informatiche.

Un altro passo che abbiamo compiuto nel nostro lavoro, è stato uno studio di fattibilità sull'utilizzo del *targeting* psicografico che ci ha permesso di verificare a tutti gli effetti come sia possibile compiere pubblicità mirata sui tratti di personalità delle persone. Anche questo elemento è stato studiato nell'ottica di applicazione all'immatricolazione nelle discipline STEM. Nello specifico abbiamo individuato una strategia, che con l'uso del social network *Facebook* e una lista utenti profilata psicologicamente, permette di

limitare il *gender gap* promuovendo la nascita di ambienti STEM più preformanti.

Tuttavia il nostro lavoro è stato soggetto a delle limitazioni. In primo luogo *Twitter Occupation Dataset*, il dataset che abbiamo utilizzato per l'analisi del linguaggio naturale, ci ha fornito esclusivamente le *Bag of words* degli utenti senza contesto. Questa limitazione, come abbiamo osservato nella generazione delle *word clouds*, non ci ha permesso di mettere in luce pienamente le correlazioni dominanti tra le parole e le polarità dei tratti di personalità *Big Five*. Inoltre il dataset è stato generato su post pubblici di *Twitter*, e come già accennato sopra, questo può includere una certa dominanza di tratti di personalità (bassa estroversione e alto nevroticismo), a discapito di altri che nella realtà invece potrebbero contraddistinguere una categoria occupazionale.
Inoltre non siamo stati in grado di escludere la possibilità che i tratti di personalità di studenti e lavoratori, anche in una stessa categoria occupazionale, siano soggetti a cambiamento (ad esempio, per qualche correlazione implicita con il periodo di carriera e lo stress generato da particolari responsabilità di mansione).

6.2 Future applicazioni

Di seguito sono discussi l'integrazione del *targeting* psicografico nei Sistemi di Raccomandazione e possibili campi d'applicazione di questo lavoro di tesi.

6.2.1 Sistemi di Raccomandazione

Il nostro lavoro si presta a essere integrato all'interno di Sistemi di Raccomandazione. Seguendo quest'ottica l'interazione con la piattaforma pubblicitaria di *Facebook*, fondamentale nella nostra ricerca per individuare un target potenziale con specifici tratti di personalità, è sostituibile con un *Personality-aware Recommender System* [85, 37] che integra al suo interno sia la profilazione psicologica che l'intervento psicologicamente informato.
Su tale *Personality-aware Recommender System*, il quale raccomanda *items* agli utenti sulla base dei tratti di personalità su loro dedotti tipicamente da questionari somministrati durante la registrazione al sistema [37], abbiamo studiato una struttura alternativa a grafi [4]; allo scopo di migliorare l'accuratezza delle raccomandazioni del sistema tradizionale sfruttando la *high-order connectivity* [48] tra gli interventi informati. Ogni grafo è una polarità dei tratti di personalità *Big Five*, e ogni nodo rappresenta uno specifico *item* (annuncio pubblicitario o altro) associato con un peso a ciascuna caratteristica del tratto; inoltre se presente un arco tra due nodi questo indica similarità tra i due *items* a cui i nodi si riferiscono. Tale sistema a grafi, quando integrato all'interno di un applicativo, permetterebbe di rilevare i tratti di personalità (con questionari o *digital footprints*) degli utenti che vi si interfacciano e di connetterli ad almeno un nodo. Su questa serie di connessioni e mediante l'utilizzo della personalità più incidente (per alleviare il fenomeno della *cold-start*) su *Item TOP-N recommendation* con peso maggiore, prevediamo sia possibile indirizzarvi gli *items* più appropriati per l'utente. Inoltre per bilanciare le raccomandazioni in base al tratto di personalità, definiamo accettabile che ogni qualvolta una raccomandazione abbia un esito negativo il peso dell'*item* associato al nodo subisca una penalizzazione e i vicini invece un incremento per un fattore costante; altrimenti è lo stesso nodo raccomandato che viene incrementato per sottolineare la relazione tra tratto di personalità e raccomandazione.

La struttura del *Personality-aware Recommender System* consentirebbe l'applicazione diretta del *targeting* psicografico senza vincoli verso un Media digitale esistente, e di svolgere conseguentemente profilazione psicologica e intervento psicologicamente informato all'interno di un unico applicativo.
Gli effetti che prevediamo di questo *Personality-aware Recommender System* applicato al caso STEM sono:

* La possibilità di utilizzare più *digital footprints* personalizzate in base al contesto;
* Non limitarsi esclusivamente ad annunci pubblicitari, ma consigliare anche particolari contenuti (approfondimenti, argomenti interessanti per la personalità);
* La portabilità del sistema in contesti differenti da STEM, senza dover modificare la sua struttura interna.

6.2.2 Campi d'applicazione

I nostri risultati si prestano a essere estesi in una molteplicità di campi e contesti. Di seguito ne discutiamo alcuni.

6.2.2.1 NOSTEM e *bias* sociali

Il tracciamento dei tratti di personalità in correlazione con le discipline NOSTEM e la ricerca di ulteriori *bias* sociali con la definizione degli interventi psicologicamente informati più adeguati.

6.2.2.2 Prevenzione di patologie e disturbi

Estensione del concetto di correlazione, non solo legata all'ambito occupazionale ma a qualsiasi scelta nella vita di un individuo. Ipotizziamo difatti che con il *targeting* psicografico si possa propriamente intervenire nella vita di un soggetto e prevenire il verificarsi di situazioni che potrebbero causare l'insorgenza di disturbi e patologie (come depressione, ansia, PTSD).

6.2.2.3 Tutela di categorie fragili e protette

La ricerca di correlazioni nei soggetti più fragili e appartenenti alle categorie protette della società (malati cronici, disabili e anziani); in modo mediante l'utilizzo del *targeting* psicografico di indirizzarvi degli interventi mirati a evitare discriminazioni, interne ed esterne all'individuo stesso. Ipotizziamo difatti che un soggetto, appartenente a queste categorie, sia più portato a ricercare un supporto da parte della società e per questo se posto in determinate condizioni avverse (come può essere un ambiente con alti livelli di nevroticismo e introversione) possa trovarsi impossibilitato a perseguire i propri obiettivi. Il *targeting* psicografico, in questo scenario, potrebbe influenzare le scelte del soggetto, in modo da evitare la rinuncia e allo stesso tempo favorire l'ambiente a essere più ospitale.

Appendice A

Twitter Occupation Dataset

Questa appendice contiene materiale addizionale al Capitolo 4.

A.1 Analisi occupazionale

Tabella A.2 mostra la categorizzazione delle occupazioni STEM, le Figure A.1, A.2, A.3 e A.4 la distribuzione delle occupazioni in *Twitter Occupation Dataset*.

STEM	
Occupation	**Job family**
Production managers and directors in manufacturing (1121)	Managers (inc. production managers)
Production managers and directors in mining and energy (1123)	Managers (inc. production managers)
Waste disposal and environmental services managers (1255)	Managers (inc. production managers)
Chemical scientists (2111)	Scientists
Biological scientists and biochemists (2112)	Scientists
Physical scientists (2113)	Scientists
Natural and social science professionals n.e.c. (2119)	Scientists
Civil engineers (2121)	Engineering professionals
Mechanical engineers (2122)	Engineering professionals
Electrical engineers (2123)	Engineering professionals
Electronics engineers (2124)	Engineering professionals
Design and development engineers (2126)	Engineering professionals
Production and process engineers (2127)	Engineering professionals
Engineering professionals n.e.c. (2129)	Engineering professionals
IT specialist managers (2133)	IT professionals
Information technology and telecommunications directors (1136)	IT professionals
IT project and programme managers (2134)	IT professionals
IT business analysts, architects and systems designers (2135)	IT professionals
Programmers and software development professionals (2136)	IT professionals
Web design and development professionals (2137)	IT professionals
Information technology and telecommunications professionals n.e.c. (2139)	IT professionals
Conservation professionals (2141)	Environment / conservation professionals
Environment professionals (2142)	Environment / conservation professionals
Research and development managers (2150)	R&D managers
Quality control and planning engineers (2461)	Quality professionals
Quality assurance and regulatory professionals (2462)	Quality professionals
Environmental health professionals (2463)	Quality professionals
Laboratory technicians (3111)	Science, engineering, production, technicians
Electrical and electronics technicians (3112)	Science, engineering, production, technicians
Engineering technicians (3113)	Science, engineering, production, technicians
Building and civil engineering technicians (3114)	Science, engineering, production, technicians
Quality assurance technicians (3115)	Science, engineering, production, technicians
Planning, process and production technicians (3116)	Science, engineering, production, technicians
Science, engineering and production technicians n.e.c. (3119)	Science, engineering, production, technicians
IT operations technicians (3131)	IT Technicians
IT user support technicians (3132)	IT Technicians
Health and safety officers (3567)	Health and safety officers
IT engineers (5245)	IT engineers
Actuaries, economists and statisticians (2425)	Business, research and administrative professionals

Tabella A.2: *Occupazioni STEM.*
Categorizzazione delle occupazioni STEM, come presentate nel lavoro *UK Commission for Employment and Skills 2015* [40]; con l'aggiunta di *SOC code* 2425.

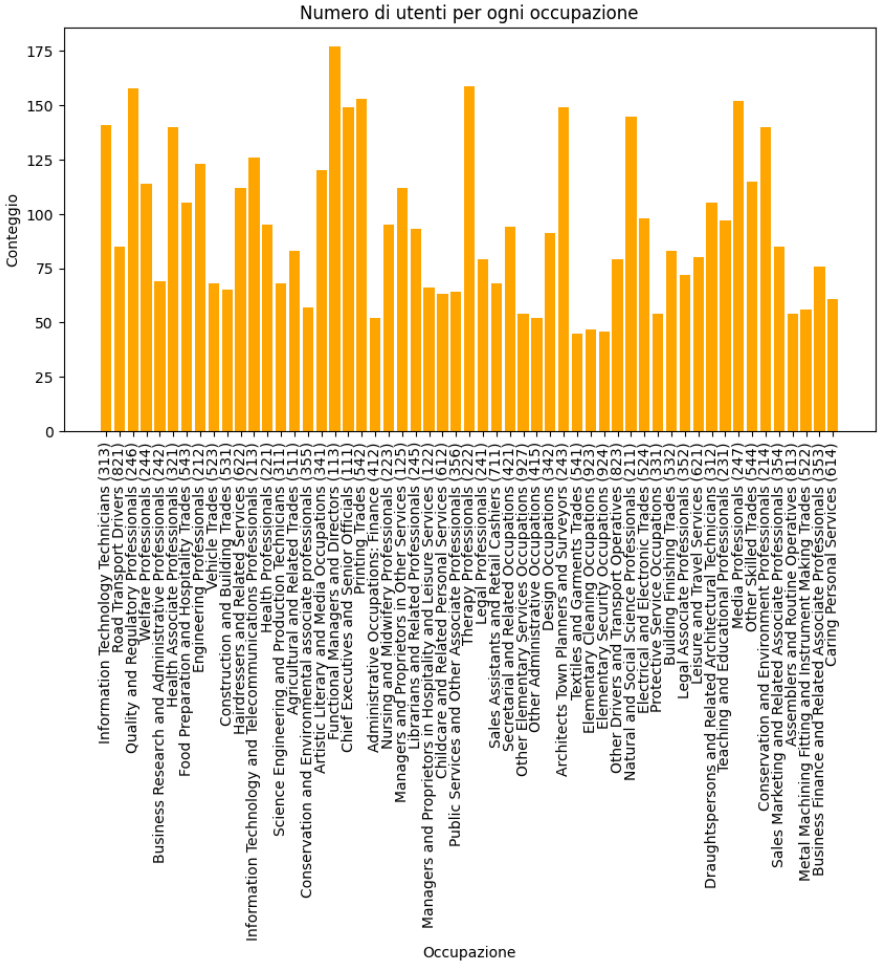


Figura A.1: *Distribuzione degli utenti sulle occupazioni - Twitter Occupation Dataset.* Il numero di utenti lavoratori analizzato è stato di 5'189.

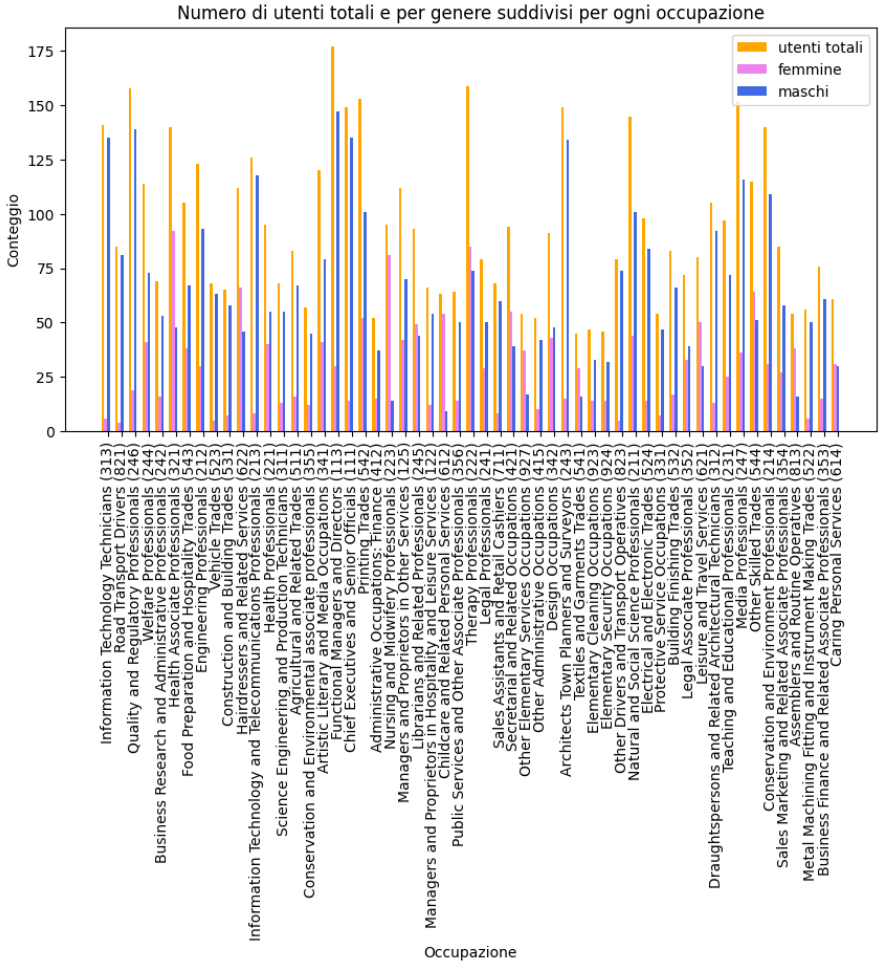


Figura A.2: *Distribuzione degli utenti in base al genere sulle occupazioni - Twitter Occupation Dataset.*
Su 5'189 lavoratori 1'612 sono risultati femmine e 3'577 maschi. Non tutte le occupazioni sono a maggioranza maschile; ne sono un esempio in questa direzione le occupazioni legate al tempo libero (*Leisure and Travel Services (621)* con 50 femmine e 30 maschi) e i servizi di assistenza (*Childcare and Related Personal Services (612)* con 54 femmine e 9 maschi).

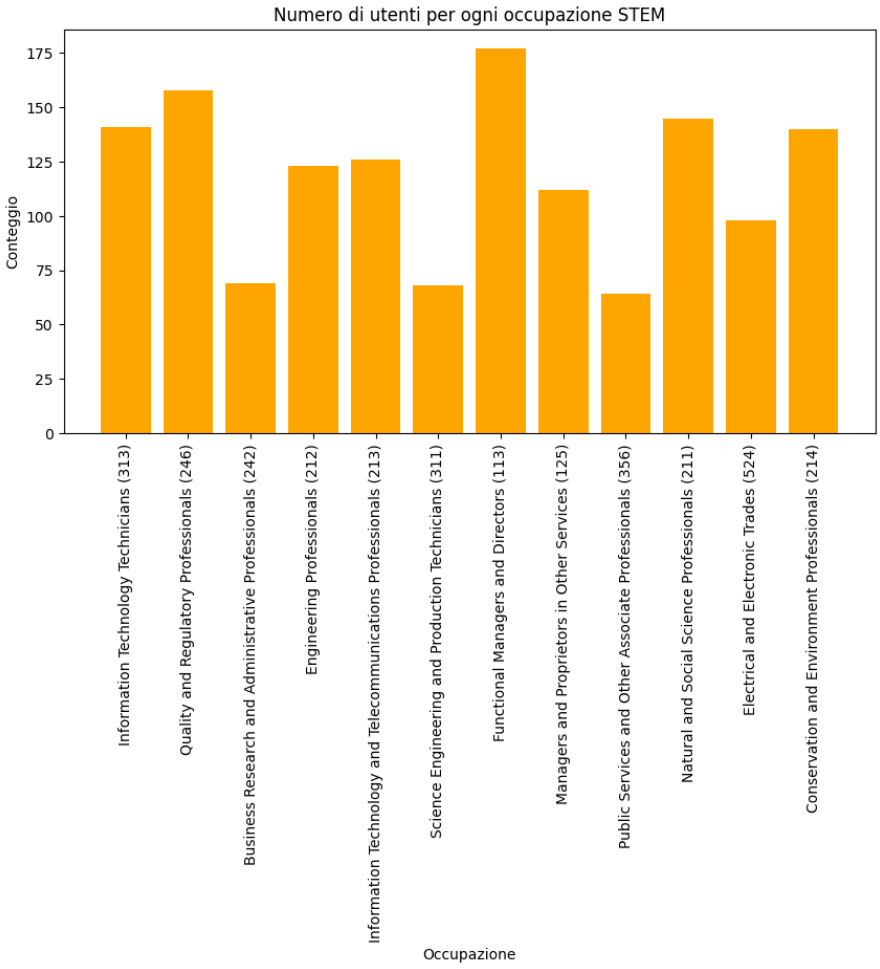


Figura A.3: *Distribuzione degli utenti sulle occupazioni STEM - Twitter Occupation Dataset.* Il totale di utenti lavoratori STEM analizzato è stato di 1'421.

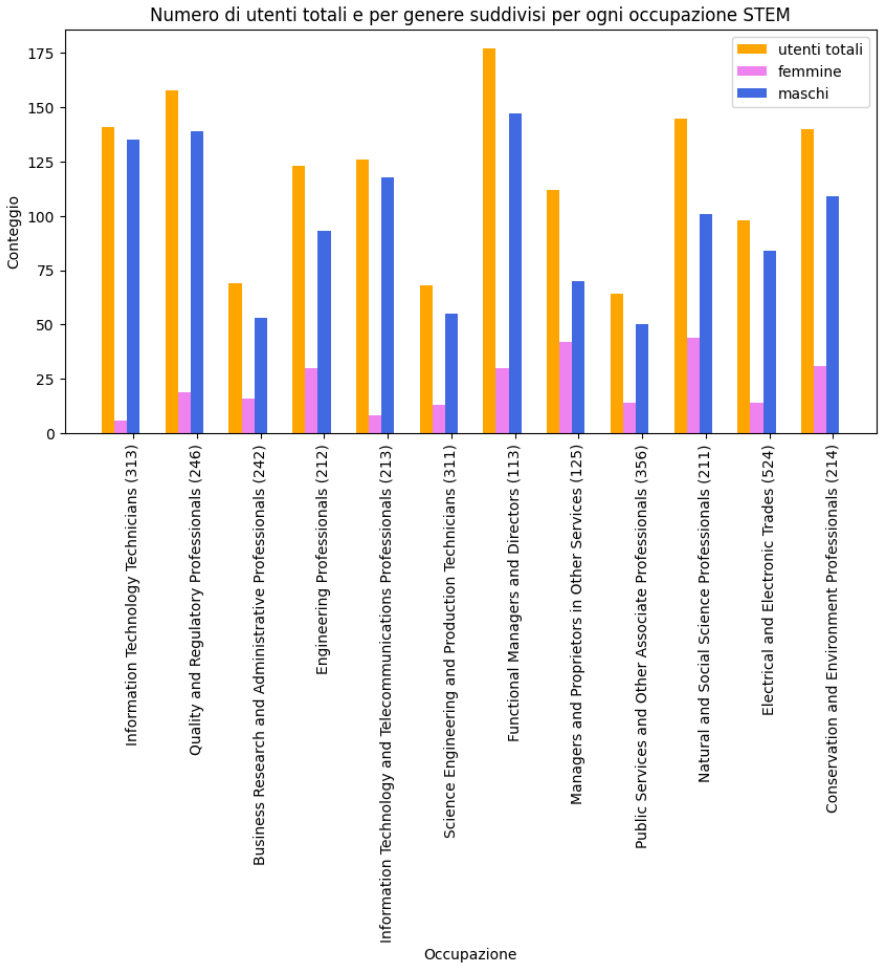


Figura A.4: *Distribuzione degli utenti in base al genere sulle occupazioni STEM - Twitter Occupation Dataset.*
Su 1'421 lavoratori 267 sono risultati femmine e 1'154 maschi. Tutte le occupazioni sono a maggioranza maschile; tuttavia le femmine preferiscono lavori STEM con un approccio sociale/umano e a utilità marcata verso terzi, come *Natural and Social Science Professionals (211) e Managers and Proprietors in Other Services (125)*; a discapito di altre occupazioni davanti al terminale, come *Information Technology and Telecommunications Professionals (213) e Information Technology Technicians (313)*. Per i maschi la tendenza è contraria.

A.2 *Word clouds*

Tabelle A.3 e A.4 mostrano le primo 100 parole femminili e maschili, che abbiamo individuato dalla generazione di *word clouds* con strategia a *clustering* per *term frequency* sulle parole di *Twitter Occupation Dataset*. Invece Figura A.5 presenta tutte le *word clouds* di personalità che abbiamo realizzato con approccio *Differential Language Analysis* applicato alle parole di *Twitter Occupation Dataset*.

Parole femminili
love, editing, xx, day, tp, today, lovely, know, hair, much, just, life, lol, feel, little, beautiful, happy, girl, baby, stories, excited, xxx, kids, wait, like, book, fun, make, amazing, go, miss, cute, get, bed, photo, birthday, dress, escort, gw, girls, things, night, sleep, time, ready, posted, gorgeous, going, friends, friend, shop, gift, heart, female, sweet, made, favorite, food, health, school, woot, fab, perfect, look, show, feeling, eat, chocolate, tomorrow, summer, tea, home, everyone, mom, mau, person, wonderful, care, fabulous, wedding, hate, loved, weekend, morning, cake, family, dear, dance, haha, shopping, help, people, makeup, dinner, deh, hahaha, vintage, dog, hot, wanna

Tabella A.3: *Le 100 parole femminili più significative.*

Parole maschili
game, mate, video, team, man, pic, years, data, patent, bro, fans, call, near, news, play, should, season, cut, money, first, top, good, great, football, site, looking, app, run, games, future, tax, mayor, business, case, players, system, public, map, web, record, player, beer, meeting, project, support, last, city, report, badge, media, website, final, shit, service, playing, government, year, update, bike, wife, big, more, match, online, played, building, wins, event, race, met, cheers, uploaded, liked, point, beat, software, facebook, free, problem, mobile, world, open, hit, unlocked, local, price, won, dude, league, security, science, deal, nice, market, shot, interesting, history, ago, million, points

Tabella A.4: *Le 100 parole maschili più significative.*

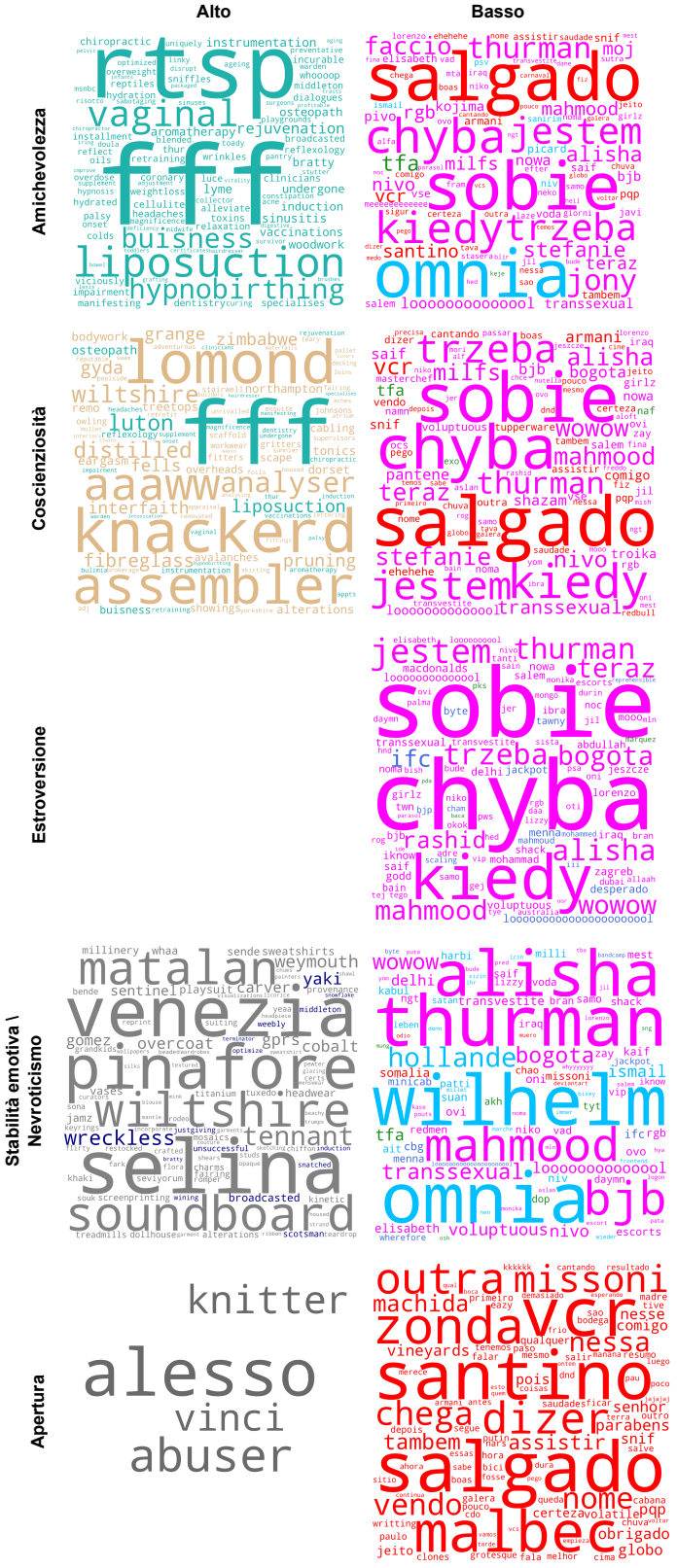


Figura A.5: *Polarità per ciascun tratto Big Five - Twitter Occupation Dataset.*

Appendice B

Pubblicità mirata online

Questa appendice contiene materiale addizionale al Capitolo 5.
Le Figure sottostanti illustrano i risultati delle indagini che abbiamo svolto sulla pubblicità mirata online e sulle modalità di tutela degli utenti dei Media digitali *Facebook, YouTube, Instagram* e *Google*.

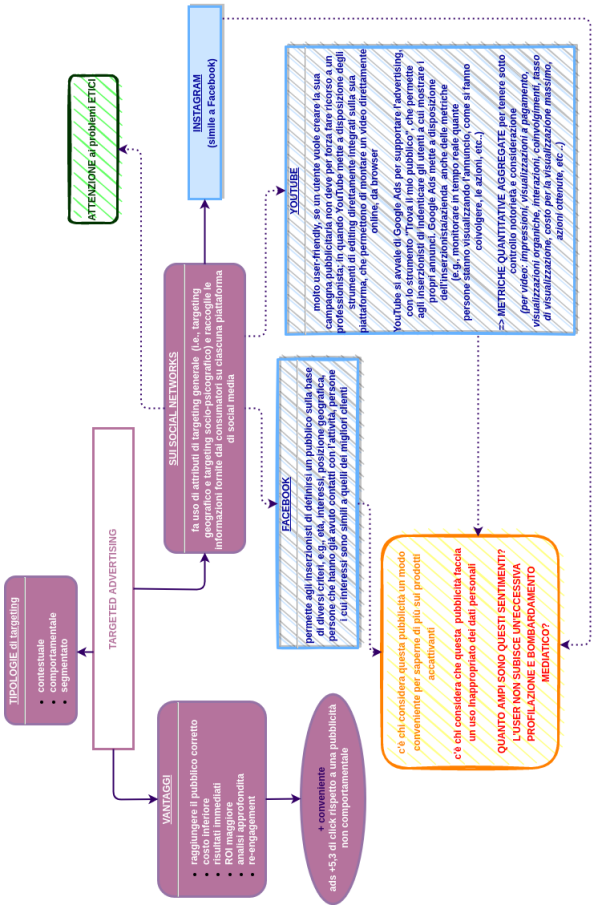


Figura B.1: *Pubblicità mirata sui social media Facebook, Instagram e YouTube.*

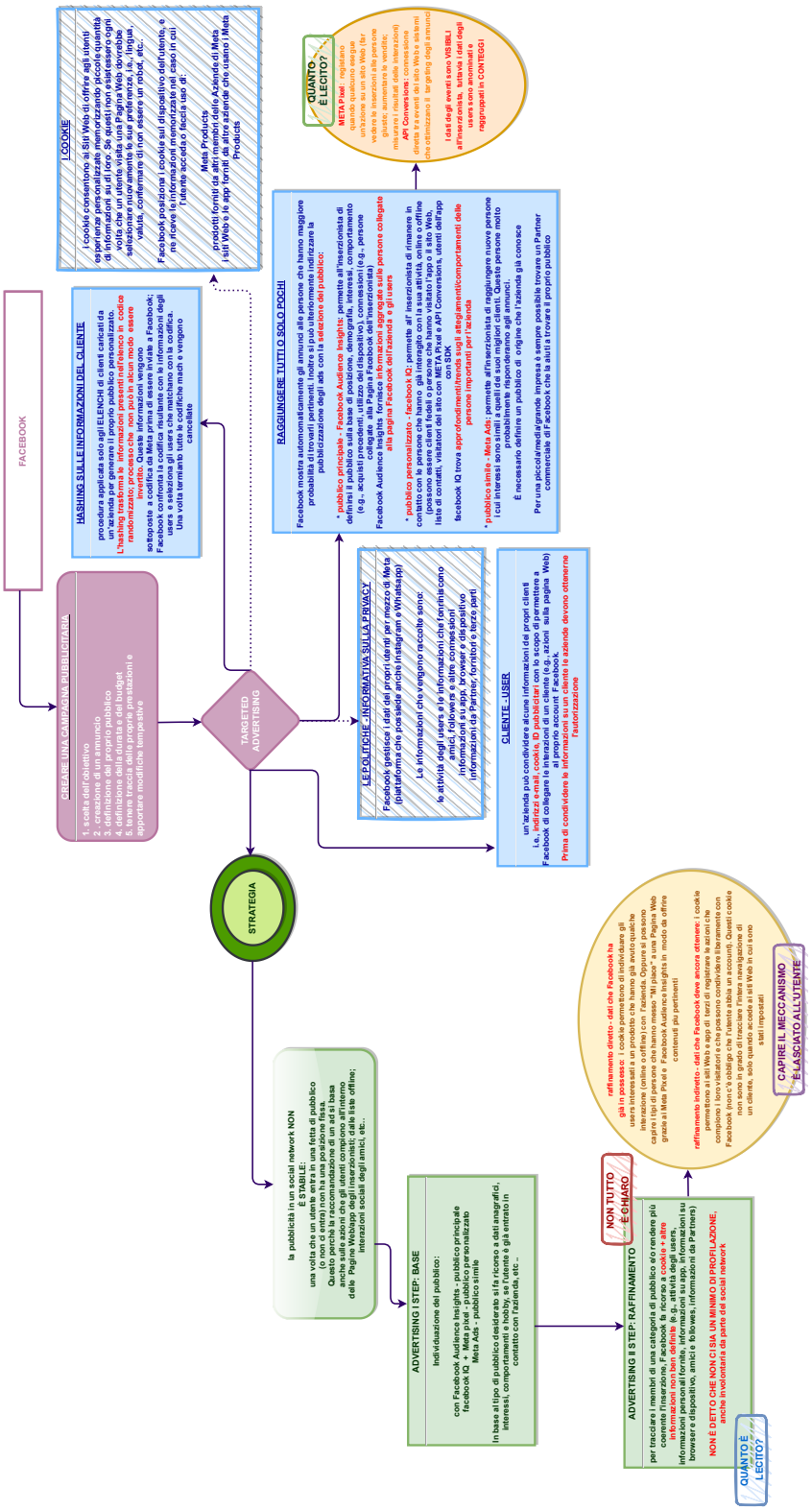


Figura B.2: *Come con Facebook si può creare una campagna pubblicitaria Meta.*

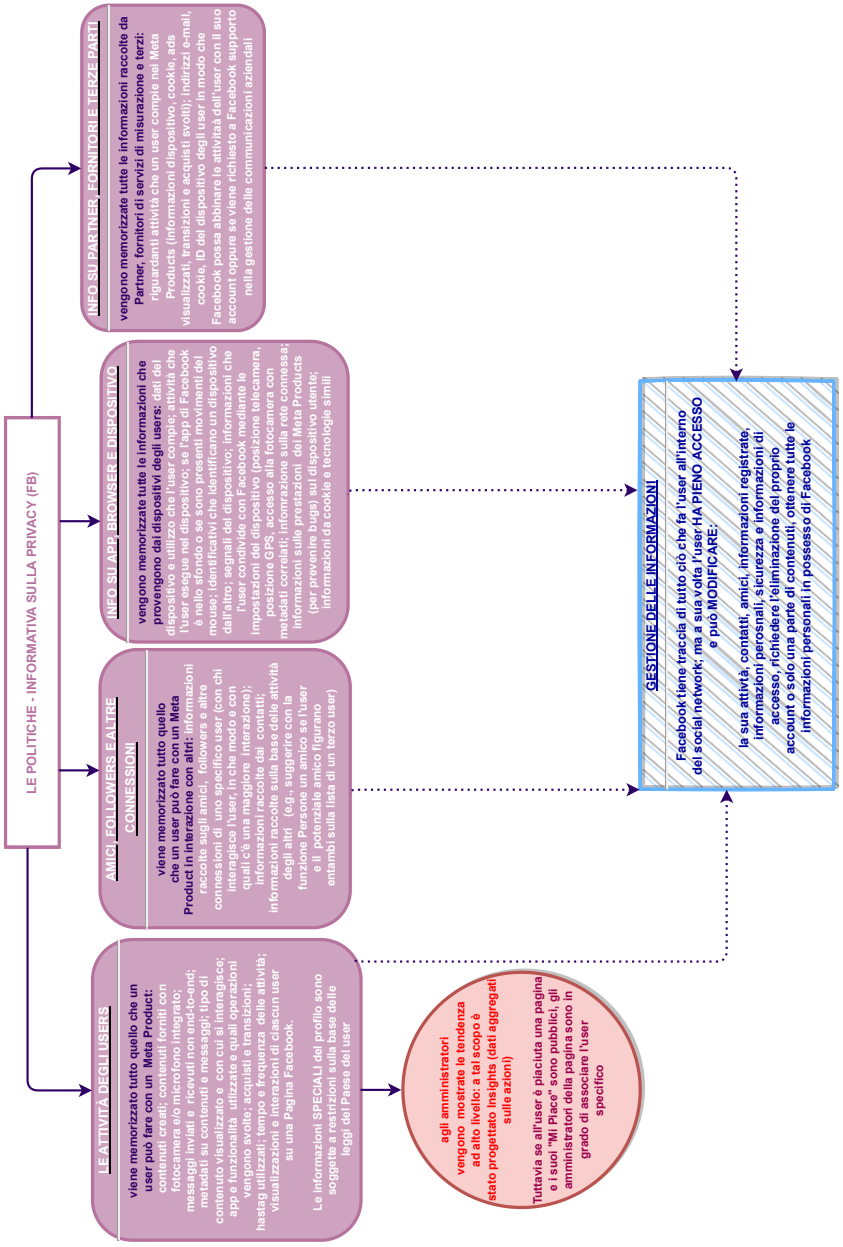


Figura B.3: *I dati che Meta raccoglie sugli utenti Facebook.*

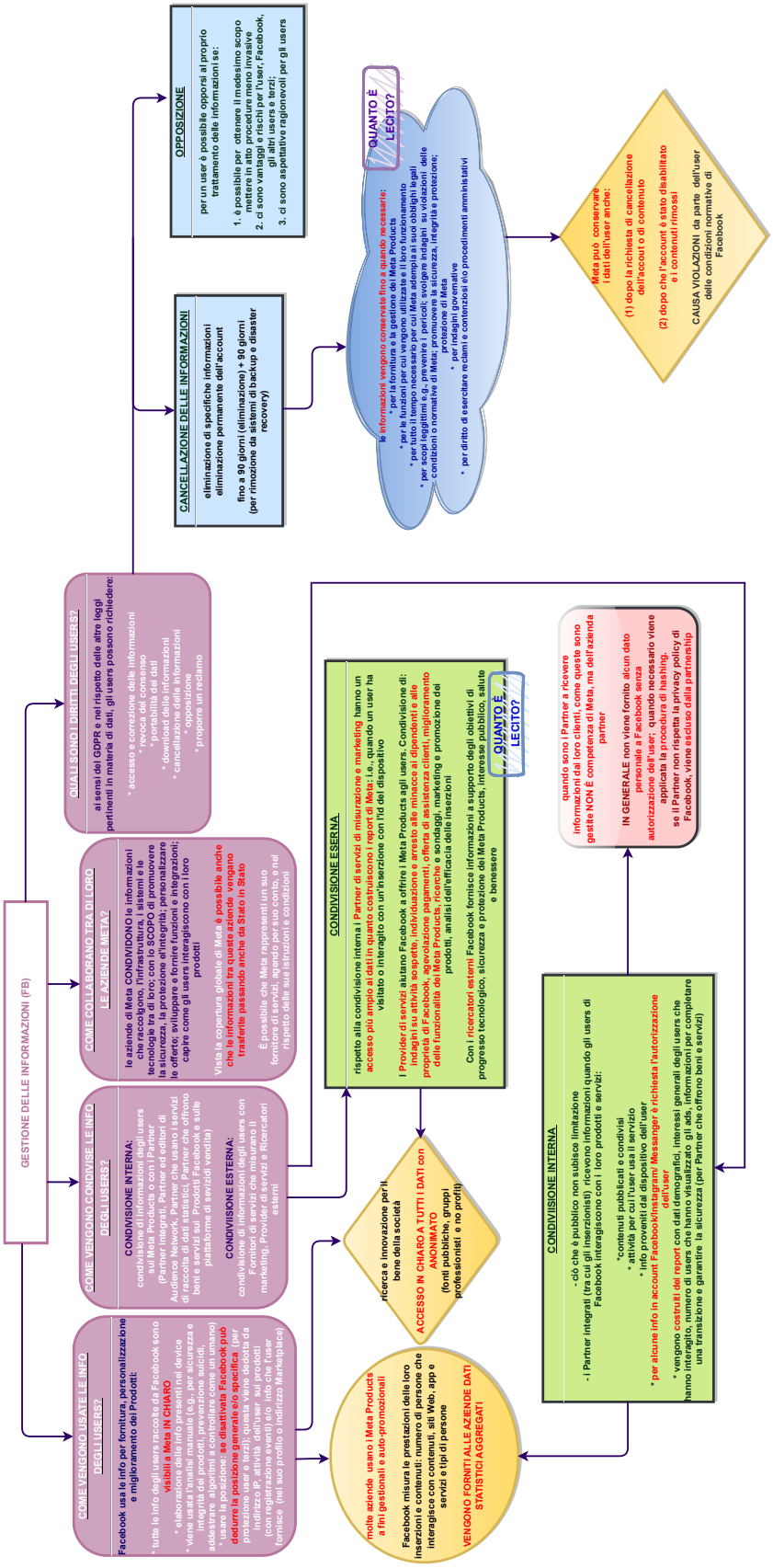


Figura B.4: *Come Meta gestisce e condivide i dati che raccoglie dagli utenti di Facebook.*

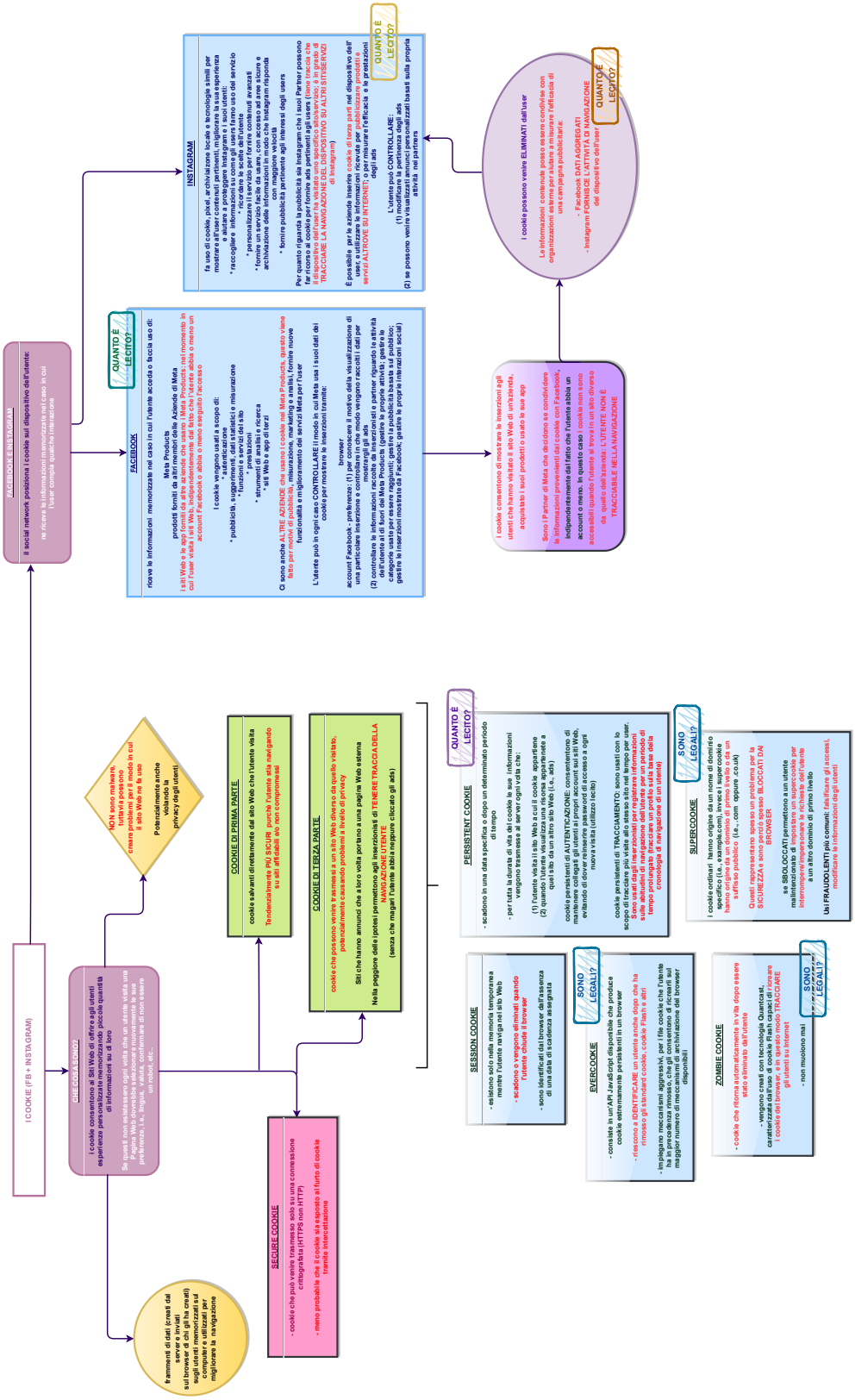


Figura B.5: *I cookie Meta in Facebook e Instagram.*

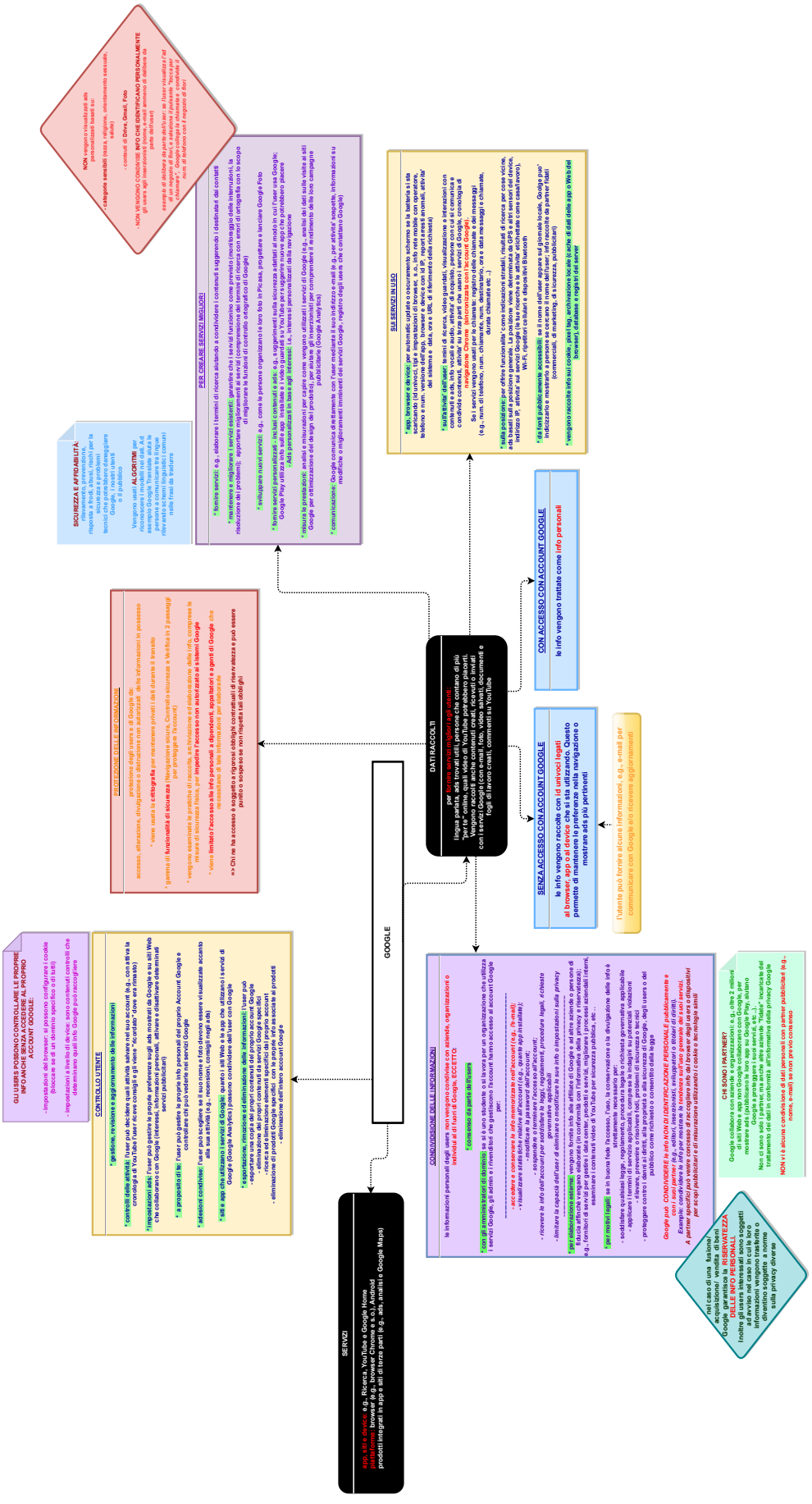


Figura B.6: *Le informazioni che raccoglie e il trattamento dei dati personali del motore di ricerca Google.*

Glossario

API *API* è l'acronimo di *Application Programming Interface*, un intermediario software che consente a due applicazioni di comunicare tra loro. Le *API* sono un modo accessibile per estrarre e condividere i dati all'interno e tra le organizzazioni (riferimento al contenuto). 66, 99

Cambridge Analytica *Cambridge Analytica* è una società di analisi dei dati che ha rivendicato un ruolo importante nella campagna *Leave*, per il referendum sull'adesione della Regno Unito all'UE. In seguito è diventata una figura chiave anche per le operazioni digitali svolte durante la campagna elettorale di Donald Trump (riferimento al contenuto). 14, 99

dati personali Nei *dati personali* rientrano tutte quelle informazioni non liberamente pubblicate dell'utente di un social networks; anche le e-mail e i "Mi Piace" di una pagina *Facebook* possono essere considerati personali, se non resi pubblicamente accessibili dall'utente che ha svolto l'azione (riferimento al contenuto). 66, 99

Display La Rete *Display* consiste in un gruppo di oltre due milioni di siti Web, video e applicazioni su cui pubblicare gli annunci pubblicitari (riferimento al contenuto). 59, 99

Intelligenza Artificiale L'*Intelligenza Artificiale* (acronimo IA) è la scienza e l'ingegneria della creazione di macchine e programmi informatici intelligenti. È legata al compito di usare i computer per comprendere l'intelligenza umana, senza le limitazioni dei metodi biologicamente osservabili (riferimento al contenuto). 10, 99

Livello 4+ Il *Livello 4+* prevede che una persona sia in possesso almeno del DipHE (Livello 5, *Diploma of Higher Education*) ed equivale ai primi due anni di un corso di laurea (riferimento al primo contenuto, riferimento al secondo contenuto). 18, 99

Machine Learning Il *Machine Learning* (acronimo ML, tradotto in Apprendimento Automatico) è una branca dell'Intelligenza Artificiale che si concentra sull'uso di dati e algoritmi per imitare, attraverso l'uso di metodi statistici e modelli, il modo in cui gli esseri umani imparano; migliorando gradualmente la loro precisione. Gli algoritmi di ML sono in genere creati utilizzando *framework* che accelerano lo sviluppo di soluzioni, come *TensorFlow* e *PyTorch* (riferimento al contenuto). 2, 99

Media digitali I *Media digitali* sono mezzi di comunicazione per la trasmissione e l'archiviazione di dati legati alle tecnologie digitali e alla rete Internet (riferimento al contenuto). 11, 100

myPersonality *myPersonality* era un'applicazione di *Facebook* che consentiva ai suoi utenti di partecipare a ricerche psicologiche compilando un questionario sulla personalità. È stata creata da David Stillwell (2007) e Michal Kosinski (2009); chiusa nel 2012 per mancanza di tempo per mantenerla (riferimento al contenuto). 20, 100

Natural Language Processing Il *Natural Language Processing* (acronimo NLP, tradotto in Elaborazione del Linguaggio Naturale) si riferisce alla branca dell'Intelligenza Artificiale che si occupa di dare ai computer la capacità di comprendere testi e parole pronunciate, in modo simile a quello che fanno gli esseri umani (riferimento al contenuto). 2, 100

Numeracy La *Numeracy* è la capacità di utilizzare o comprendere le tecniche numeriche della matematica (riferimento al contenuto). 19, 100

Office for National Statistics (ONS) L'*Office for National Statistics* è il più grande produttore indipendente di statistiche ufficiali del Regno Unito e l'istituto statistico nazionale riconosciuto del Regno Unito (riferimento al contenuto). 17, 100

Problem Solving Il *Problem Solving* è la capacità di risolvere problemi (riferimento al contenuto). 19, 100

Quoziente Intellettivo (QI) Il *Quoziente Intellettivo* è un tipo di punteggio standard che indica quanto al di sopra o al di sotto del suo gruppo di pari un individuo si trova nelle capacità mentali. Il punteggio del gruppo di pari è un QI di 100; questo si ottiene applicando lo stesso test a un numero enorme di persone di tutti gli strati socio-economici della società e facendone la media (riferimento al contenuto). 23, 100

stereotipi Uno *stereotipo* (dal greco stereòs "rigido" e tùpos "impronta") è una rappresentazione mentale o un'idea riguardo una determinata realtà. Esistono due tipologie di stereotipi: cognitivi, semplificazione delle informazioni che l'individuo immagazzina prima di entrare a far parte del suo patrimonio culturale; e sociali, immagini mentali condivise da intere società e che riguardano ampie categorie (come cristiani, musulmani, omosessuali, comunisti) (riferimento al contenuto). 1, 100

UK Commission for Employment and Skills La *UK Commission for Employment and Skills* è un'organizzazione finanziata con fondi pubblici e guidata dall'industria; e che offre indicazioni su competenze e problemi occupazionali nel Regno Unito. Alla data odierna (11/04/2023) risulta chiusa (riferimento al contenuto). 18, 100

Riferimenti

[1] Australian Human Rights Commission 2020. «Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias». In: (2020). URL: https://humanrights.gov.au/our-work/rights-and-freedoms/publications/using-artificial-intelligence-make-decisions-addressing (cit. a p. 14).

[2] Schwartz H. A. e Gomez F. «Acquiring knowledge from the web to be used as selectors for noun sense disambiguation». In: (2008), pp. 105–112 (cit. a p. 8).

[3] Sami Abu-El-Haija et al. «YouTube-8M: A Large-Scale Video Classification Benchmark». In: *CoRR* abs/1609.08675 (2016). arXiv: 1609.08675. URL: http://arxiv.org/abs/1609.08675 (cit. a p. 66).

[4] Ifeoma Adaji et al. «Personality Based Recipe Recommendation Using Recipe Network Graphs». In: *Social Computing and Social Media. Technologies and Analytics*. A cura di Gabriele Meiselwitz. Cham: Springer International Publishing, 2018, pp. 161–170 (cit. a p. 82).

[5] Google Ads. «Informazioni sui rapporti sul coinvolgimento». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/6156146?sjid=1931147236026941453-EU (cit. a p. 64).

[6] Google Ads. «Informazioni sull'impatto del brand». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/9049825?hl=it (cit. a p. 64).

[7] Google Ads. «Informazioni sulla freq. media impr. per utente (7 o 30 giorni)». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/9507337?sjid=1931147236026941453-EU (cit. a p. 64).

[8] Google Ads. «Informazioni sulle metriche relative agli annunci e alle visualizzazioni di YouTube». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/2375431 (cit. a p. 64).

[9] Google Ads. «Unique Reach: definizione». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/9012727?sjid=1931147236026941453-EU (cit. a p. 64).

[10] Azeem Akbar, Amara Malik e Nosheen Fatima Warraich. «Big Five Personality Traits and Knowledge Sharing Intentions of Academic Librarians». In: *The Journal of Academic Librarianship* 49.2 (2023), p. 102632. ISSN: 0099-1333. DOI: https://doi.org/10.1016/j.acalib.2022.102632. URL: https:

//www.sciencedirect.com/science/article/pii/S0099133322001483 (cit. alle pp. 24, 38, 39).

[11] Lina Aldén e Emma Neuman. «Culture and the gender gap in choice of major: An analysis using sibling comparisons». In: *Journal of Economic Behavior & Organization* 201 (2022), pp. 346–373. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2022.07.026. URL: https://www.sciencedirect.com/science/article/pii/S0167268122002608 (cit. a p. 1).

[12] Neal Andrew et al. «Predicting the form and direction of work role performance from the Big 5 model of personality traits». In: *Journal of Organizational Behavior* 33.2 (2011), pp. 175–192. DOI: https://doi.org/10.1002/job.742. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.742. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/job.742 (cit. a p. 76).

[13] Ruth E. Appel e Sandra C. Matz. «Chapter 6 - Psychological targeting in the age of Big Data». In: (2021). A cura di Dustin Wood et al., pp. 193–222. DOI: https://doi.org/10.1016/B978-0-12-819200-9.00015-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128192009000156 (cit. alle pp. 11–15).

[14] Danny Azucar, Davide Marengo e Michele Settanni. «Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis». In: *Personality and Individual Differences* 124 (2018), pp. 150–159. ISSN: 0191-8869. DOI: https://doi.org/10.1016/j.paid.2017.12.018. URL: https://www.sciencedirect.com/science/article/pii/S0191886917307328 (cit. a p. 11).

[15] Cattell R. B. «Confirmation and clarification of primary personality factors». In: *Psychometrika* 12 (1947), pp. 197–220 (cit. a p. 9).

[16] M.W. Kreuter B.K. Rimer. «Advancing tailored health communication: A persuasion and message effects perspective». In: *Journal of Communication* 56 (2006), S184–S201. URL: 10.1111/j.1460-2466.2006.00289.x (cit. a p. 11).

[17] Bachner-Melman et al. «Addressing the imbalance: The downside of extraversion and the upside of introversion». In: (gen. 2014), pp. 158–165 (cit. a p. 76).

[18] M. Begale et al. «Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia». In: *Schizophrenia Bulletin* 40.6 (2014), pp. 1244–1253. DOI: 10.1093/schbul/sbu033 (cit. a p. 14).

[19] David M. Blei, Andrew Y. Ng e Michael I. Jordan. «Latent Dirichlet Allocation». In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435. URL: https://dl.acm.org/doi/10.5555/944919.944937 (cit. alle pp. 7, 21).

[20] Stephen Blyth. «Karl Pearson and the Correlation Curve». In: *International Statistical Review / Revue Internationale de Statistique* 62.3 (1994), pp. 393–403. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403769 (visitato il 20/04/2023) (cit. a p. 52).

[21] Jonny Brooks-Bartlett. «Probability concepts explained: Maximum likelihood estimation». In: *Towards Data Science* (2018). Ultima visita il 06/04/2023. URL:

https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1 (cit. a p. 18).

[22] Deerwester S. C. et al. «Computer information retrieval using latent semantic structure». In: (1988) (cit. a p. 7).

[23] Deerwester S. C. et al. «Indexing by latent semantic analysis». In: 41 (1990), pp. 391–407. DOI: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AIDASI1>3.0.CO;2-9 (cit. a p. 7).

[24] Gian Vittorio Caprara e Accursio Gennaro. «Psicologia della personalità». In: Strumenti (1994). A cura di Il Mulino (cit. a p. 8).

[25] Alexander Chernev, Ulf Böckenholt e Joseph Goodman. «Choice overload: A conceptual review and meta-analysis». In: *Journal of Consumer Psychology* 25.2 (2015), pp. 333–358. ISSN: 1057-7408. DOI: https://doi.org/10.1016/j.jcps.2014.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1057740814000916 (cit. a p. 13).

[26] Nian Shong Chok. «Pearson's versus Spearman's and kendall's correlation. Coefficients for continuous data». In: (2010). URL: https://d-scholarship.pitt.edu/8056/1/Chokns_etd2010.pdf (cit. a p. 52).

[27] Deborah A. Cobb-Clark e Michelle Tan. «Noncognitive skills, occupational attainment, and relative wages». In: *Labour Economics* 18.1 (2011), pp. 1–13. ISSN: 0927-5371. DOI: https://doi.org/10.1016/j.labeco.2010.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0927537110000904 (cit. alle pp. 23, 24, 54, 57).

[28] Johan Coenen, Lex Borghans e Ron Diris. «Personality traits, preferences and educational choices: A focus on STEM». In: *Journal of Economic Psychology* 84 (2021), p. 102361. ISSN: 0167-4870. DOI: https://doi.org/10.1016/j.joep.2021.102361. URL: https://www.sciencedirect.com/science/article/pii/S0167487021000015 (cit. alle pp. 13, 23, 54, 57).

[29] European Commission, Directorate-General for Justice e Consumers. «2019 report on equality between women and men in the EU». In: (2019). DOI: doi/10.2838/395144 (cit. a p. 1).

[30] Parlamento europeo e del Consiglio. «Regolamento generale sulla protezione dei dati». In: (2016). Visitato il 29/04/2023. URL: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ITA&toc=OJ:L:2016:119:TOC (cit. alle pp. 15, 67, 68).

[31] Digital Marketing Consulting. «Come Fare Una Campagna Google ADS - Guida Pratica Per Principianti». In: (2022). Visualizzato il 29/04/2023. URL: https://www.youtube.com/watch?v=fUVpkf_Rb38 (cit. alle pp. 63, 65).

[32] Christianne Corbett e Catherine Hill. «Solving the Equation: The Variables for Women's Success in Engineering and Computing». In: (mar. 2015). URL: https://www.aauw.org/app/uploads/2020/03/Solving-the-Equation-report-nsa.pdf (cit. alle pp. 1, 2, 79).

[33] Burger John D. et al. «Discriminating Gender on Twitter». In: (2011) (cit. a p. 8).

[34] Bamman David, Jacob Eisenstein e Schnoebelen Tyler. «Gender identity and lexical variation in social media». In: *Journal of Sociolinguistics* 18.2 (2014), pp. 135–160. DOI: https://doi.org/10.1111/josl.12080. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12080. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/josl.12080 (cit. alle pp. 40–42).

[35] F. N. David e C. L. Mallows. «The Variance of Spearman's Rho in Normal Samples». In: *Biometrika* 48.1/2 (1961), pp. 19–28. ISSN: 00063444. URL: http://www.jstor.org/stable/2333126 (visitato il 20/04/2023) (cit. a p. 52).

[36] John DeJesus. «Point Biserial Correlation with Python». In: (2019). URL: https://towardsdatascience.com/point-biserial-correlation-with-python-f7cd591bd3b1 (cit. a p. 53).

[37] Sahraoui Dhelim et al. «A Survey on Personality-Aware Recommendation Systems». In: *CoRR* abs/2101.12153 (2021). arXiv: 2101.12153. URL: https://arxiv.org/abs/2101.12153 (cit. a p. 82).

[38] J. M. Digman. «Personality structure: Emergence of the five-factor model». In: *Annual Review of Psychology* 41 (1990), pp. 417–440 (cit. a p. 9).

[39] Francis M. E. e Pennebaker J. W. «LIWC: Linguistic inquiry and word count». In: (1993) (cit. a p. 6).

[40] UK Commission for Employment e Skills (UKCES). «Reviewing the requirement for high level STEM skills». In: (2015). viewed 31 Mar 2023. URL: https://www.gov.uk/government/publications/high-level-stem-skills-requirements-in-the-uk-labour-market (cit. alle pp. 18, 19, 48, 85).

[41] Piero Esposito e Sergio Scicchitano. «Drivers of skill mismatch among Italian graduates: the role of personality traits». In: *Applied Economics* 0.0 (2022), pp. 1–22. DOI: 10.1080/00036846.2022.2130151. eprint: https://doi.org/10.1080/00036846.2022.2130151. URL: https://doi.org/10.1080/00036846.2022.2130151 (cit. alle pp. 25, 26, 37).

[42] «Fai Crescere il tuo Video su YouTube!» In: (2023). Ultima visita il 26/04/2023. URL: https://easyadv.co/acquistare-visualizzazioni-youtube/ (cit. a p. 63).

[43] D. W. Fiske. «Consistency of the factorial structures of personality ratings from different sources». In: *Journal of Abnormal and Social Psychology* 44.3 (1949), pp. 329–344 (cit. a p. 9).

[44] Fleur Förster. «How to reach your target group more individually with psychographic targeting». In: *dmexco* (2020). URL: https://dmexco.com/stories/how-to-reach-your-target-group-more-individually-with-psychographic-targeting/ (cit. a p. 11).

[45] Niklas FOURBERG et al. «Online advertising: the impact of targeted advertising on advertisers, market access and consumer choice». In: (2021). URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662913/IPOL_STU(2021)662913_EN.pdf (cit. a p. 59).

[46] Kendall M. G. «A new measure of rank correlation». In: *Biometrika* 30.1-2 (giu. 1938), pp. 81–93. ISSN: 0006-3444. DOI: 10.1093/biomet/30.1-2.81. eprint:

https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf. URL: https://doi.org/10.1093/biomet/30.1-2.81 (cit. a p. 52).

[47] Kathryn Galanis, Ishana Syed e Kat Williams. «Privacy versus Products in Targeted Digital Advertising». In: *Media Ethics, Initiative Center for Media Engagement University of Texas at Austin* (2021) (cit. a p. 60).

[48] Chen Gao et al. «Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions». In: *CoRR* abs/2109.12843 (2021). arXiv: 2109.12843. URL: https://arxiv.org/abs/2109.12843 (cit. a p. 82).

[49] Guido Gili. «Il problema della manipolazione: peccato originale dei media?» In: (2001) (cit. a p. 14).

[50] K. Glanz e J.E. Stryker. «Health Behavior and Risk Factors». In: (2008). A cura di Harald Kristian (Kris) Heggenhougen, pp. 146–152. DOI: https://doi.org/10.1016/B978-012373960-5.00099-X. URL: https://www.sciencedirect.com/science/article/pii/B978012373960500099X (cit. a p. 11).

[51] L. R. Goldberg. «The structure of phenotypic personality traits». In: *American Psychologist* 48 (1993), pp. 26–34. URL: http://psych.colorado.edu/~carey/courses/psyc5112/readings/psnstructure_goldberg.pdf (cit. a p. 9).

[52] Google. «Ampliare le possibilità con Google Cloud Platform». In: (2023). URL: https://cloud.google.com/edu/researchers (cit. a p. 67).

[53] Google. «Collaborations with the research and academic communities». In: (2023). URL: https://research.google/outreach/ (cit. a p. 67).

[54] Google. «Fai crescere la tua attività con Google Ads». In: (2023). Ultima visita il 29/04/2023. URL: https://ads.google.com/intl/it_it/home/ (cit. alle pp. 62, 65).

[55] Goolge. «Norme sulla privacy». In: (2022). Ultima visita il 29/04/2023. URL: https://policies.google.com/privacy?hl=it (cit. a p. 66).

[56] Carla J Groom e James W Pennebaker. «Words». In: *Journal of Research in Personality* 36.6 (2002), pp. 615–621. ISSN: 0092-6566. DOI: https://doi.org/10.1016/S0092-6566(02)00512-3. URL: https://www.sciencedirect.com/science/article/pii/S0092656602005123 (cit. a p. 6).

[57] Amy Guttman. «Set To Take Over Tech: 70% Of Iran's Science And Engineering Students Are Women». In: (2015). URL: https://slate.com/human-interest/2017/11/the-stem-paradox-why-are-muslim-majority-countries-producing-so-many-female-engineers.html (cit. a p. 2).

[58] J. Harris. «Word clouds considered harmful». In: (2011). URL: https://www.niemanlab.org/2011/10/word-clouds-considered-harmful/ (cit. alle pp. 22, 40).

[59] J.R. Hauser, G. Liberali G.L. Urban e M. Braun. «Website morphing». In: *Marketing Science* 28.2 (2009), pp. 202–223. URL: 10.1287/mksc.1080.0459 (cit. a p. 12).

[60] J.B. Hirsh e G.V. Bodenhausen S.K. Kang. «Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits». In: *Psychological Science* 23.6 (2012), pp. 578–581. URL: 10.1177/0956797611436349 (cit. a p. 11).

[61] IBM. «Watson personality insights [online]». In: (2018). Vistato il 21/04/2023. URL: https://www.ibm.com/cloud/watson-natural-language-understanding?mhsrc=ibmsearch_a&mhq=Watson%5C%20Personality%5C%20Insights (cit. alle pp. 10, 13).

[62] ILOSTAT. «ILO Survey Catalogue». In: (2023). Visitato nei mesi di gennaio e febbraio 2023. URL: https://www.ilo.org/surveyLib/index.php/collections/LFS (cit. a p. 19).

[63] «Informazioni sulla visibilità e sulle metriche usate nei report sulla Visualizzazione attiva». In: (2023). Ultima visita il 27/04/2023. URL: https://support.google.com/google-ads/answer/7029393?sjid=1931147236026941453-EU (cit. a p. 64).

[64] InternetLiveStats. «Google Search Statistics». In: (2023). Ultima visita il 28/04/2023. URL: https://www.internetlivestats.com/google-search-statistics/ (cit. a p. 65).

[65] Ishwarappa e J. Anuradha. «A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology». In: *Procedia Computer Science* 48 (2015). International Conference on Computer, Communication and Convergence (ICCC 2015), pp. 319–324. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.04.188. URL: https://www.sciencedirect.com/science/article/pii/S1877050915006973 (cit. a p. 12).

[66] ISTAT. «Gli stereotipi sui ruoli di genere e l'immagine sociale della violenza sessuale». In: (2019). URL: https://www.istat.it/it/archivio/235994 (cit. a p. 76).

[67] Stone P. J., Dunphy D. C. e Smith M. S. «The general inquirer: A computer approach to content analysis». In: (1966) (cit. a p. 6).

[68] Eichstaedt JC et al. «Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations». In: 26.4 (2021), pp. 398–427. DOI: 10.1037/met0000349 (cit. alle pp. 6–8).

[69] O. P. John, Donahue E. M. e Kentle R. L. «The Big Five Inventory-Versions 4a And 54». In: (1991) (cit. a p. 10).

[70] Oliver P. John e Sanjay Srivastava. «The Big Five Trait taxonomy: History, measurement, and theoretical perspectives.» In: (1999) (cit. alle pp. 9, 45).

[71] Michael I. Jordan. «Why the logistic function? A tutorial discussion on probabilities and neural networks». In: (1995). URL: https://www.ics.uci.edu/~smyth/courses/cs274/readings/jordan_logistic.pdf (cit. a p. 18).

[72] A. F. Jorm. «Sex Differences in Neuroticism: A Quantitative Synthesis of Published Research». In: *Australian & New Zealand Journal of Psychiatry* 21.4 (1987). PMID: 3329513, pp. 501–506. DOI: 10.3109/00048678709158917. eprint: https://doi.org/10.3109/00048678709158917. URL: https://doi.org/10.3109/00048678709158917 (cit. alle pp. 25, 57).

[73] Machine learning journey. «K-Means And K-Medians». In: *Machine learning journey* (2023). Vistato il 06/04/2023. URL: https://machinelearningjourney.com/index.php/2020/02/07/k-means-k-medians/ (cit. a p. 19).

[74] P.T. Costa Jr. e R.R. McCrae. «NEO PI-R: Professional manual Odessa». In: *Psychological Assessment Resources (USA)* (1992) (cit. a p. 10).

[75] Rachubińska K. et al. «The relationship between women's personality traits and addiction to social networking sites on the example of Facebook. European review for medical and pharmacological sciences». In: 26.6 (2022), pp. 1809–1815. URL: https://doi.org/10.26355/eurrev_202203_28324 (cit. alle pp. 25, 54, 57).

[76] Margaret L. Kern et al. «The Online Social Self: An Open Vocabulary Approach to Personality». In: *Assessment* 21.2 (2014). PMID: 24322010, pp. 158–169. DOI: 10.1177/1073191113514104. eprint: https://doi.org/10.1177/1073191113514104. URL: https://doi.org/10.1177/1073191113514104 (cit. alle pp. 10, 13, 20–22, 34, 39, 40, 42–47, 53).

[77] «Pearson's Correlation Coefficient». In: (2008). A cura di Wilhelm Kirch, pp. 1090–1091. DOI: 10.1007/978-1-4020-5614-7_2569. URL: https://doi.org/10.1007/978-1-4020-5614-7_2569 (cit. a p. 52).

[78] Vivek Kulkarni et al. «Latent human traits in the language of social media: An open-vocabulary approach». In: *PLOS ONE* 13.11 (nov. 2018), pp. 1–18. DOI: 10.1371/journal.pone.0201703. URL: https://doi.org/10.1371/journal.pone.0201703 (cit. a p. 11).

[79] S. LaRochelle-Côté e D. W. Hango. «Overqualification, skills and job satisfaction. Insights on Canadian Society». In: (2016). URL: https://files.eric.ed.gov/fulltext/ED585327.pdf (cit. alle pp. 19, 37).

[80] Fidelia Law et al. «Children's Gender Stereotypes in STEM Following a One-Shot Growth Mindset Intervention in a Science Museum». In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.641695. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.641695 (cit. a p. 1).

[81] Mark Leary. «Neuroticism: The Big Five Personality Types Explained». In: (2018). URL: https://www.wondriumdaily.com/neuroticism-big-five-personality-types-explained/ (cit. a p. 76).

[82] Vincent LeBlanc e Michael Cox. «Interpretation of the point-biserial correlation coefficient in the context of a school examination». In: *The Quantitative Methods for Psychology* 13 (gen. 2017), pp. 46–56. DOI: 10.20982/tqmp.13.1.p046 (cit. a p. 53).

[83] Jae Sik Lee e Jin Chun Lee. «Music for My Mood: A Music Recommendation System Based on Context Reasoning». In: (2006). A cura di Paul Havinga et al., pp. 190–203 (cit. a p. 12).

[84] Sarah-Jane Leslie et al. «Expectations of brilliance underlie gender distributions across academic disciplines». In: *Science* 347.6219 (2015), pp. 262–265. DOI: 10.1126/science.1261375. eprint: https://www.science.org/doi/pdf/

10.1126/science.1261375. URL: https://www.science.org/doi/abs/10.1126/science.1261375 (cit. a p. 1).

[85] Elisabeth Lex et al. «Psychology-informed Recommender Systems». In: (2021). URL: https://elisabethlex.info/docs/2021fntir-psychology.pdf (cit. a p. 82).

[86] Pandas library. «pandas.DataFrame.corr API». In: (2023). Visitato nei mesi di gennaio - marzo. URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html (cit. a p. 48).

[87] Maristella Lunardon, Tania Cerni e Raffaella I. Rumiati. «Numeracy Gender Gap in STEM Higher Education: The Role of Neuroticism and Math Anxiety». In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.856405. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.856405 (cit. alle pp. 23, 24, 53, 54).

[88] Meyer M., Cimpian A. e Leslie S. J. «Women are underrepresented in fields where success is believed to require brilliance». In: *Frontiers in psychology* 6.235 (2015). URL: https://doi.org/10.3389/fpsyg.2015.00235 (cit. a p. 1).

[89] Peters M. et al. «Dissecting contextual word embeddings: Architecture and representation». In: (2018), pp. 1499–1509 (cit. a p. 8).

[90] Schueller Stephen M., Aguilera Adrian e David C. Mohr. «Ecological momentary interventions for depression and anxiety». In: *Depression and Anxiety* 34.6 (2017), pp. 540–545. DOI: https://doi.org/10.1002/da.22649. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22649. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/da.22649 (cit. alle pp. 12, 14).

[91] Shumanov M., Cooper H. e Ewing M. «"Using AI predicted personality to enhance advertising effectiveness"». In: *European Journal of Marketing* 56.6 (2022), pp. 1590–1609. URL: https://doi.org/10.1108/EJM-12-2019-0941 (cit. alle pp. 2, 10, 11, 13).

[92] Susie Marino. «165 Strategy-Changing Digital Marketing Statistics for 2023». In: (2023). Visitato il 01/05/2023. URL: https://www.wordstream.com/blog/ws/2022/04/19/digital-marketing-statistics (cit. alle pp. 60, 71).

[93] S.C. Matz, G. Nave M. Kosinski e D.J. Stillwell. «Psychological targeting as an effective approach to digital mass persuasion». In: *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12714–12719. URL: 10.1073/pnas.1710966114 (cit. a p. 12).

[94] Andrew Kachites McCallum. «MALLET: A Machine Learning for Language Toolkit». In: (2002). Ultima visita nel mese di marzo 2023. URL: http://mallet.cs.umass.edu (cit. a p. 21).

[95] John McCarthy. «What is Artificial Intelligence?» In: (2004) (cit. a p. 10).

[96] Stacey McLachlan. «85+ Important Social Media Advertising Statistics to Know». In: (2023). Visitato il 01/05/2023. URL: https://blog.hootsuite.com/social-media-advertising-stats/ (cit. a p. 60).

[97] Meta. «Data for Good at Meta (previously Facebook)». In: (2019). Ultima visita il 28/04/2023. URL: https://data.humdata.org/organization/facebook (cit. a p. 66).

[98] Meta. «I dati in tempo reale migliorano la nostra risposta alle crisi del mondo reale». In: (2023). Ultima visita il 28/04/2023. URL: https://dataforgood.facebook.com/dfg/about (cit. a p. 66).

[99] Meta. «Meta Business Suite». In: (2023). URL: https://www.facebook.com/business/tools/meta-business-suite (cit. a p. 60).

[100] Meta. «Rendere i dati disponibili per la ricerca indipendente». In: (2023). Ultima visita il 28/04/2023. URL: https://research.facebook.com/data/ (cit. a p. 66).

[101] Meta. «Targetizzazione delle inserzioni per il pubblico». In: (2023). Ultima visita il 26/04/2023. URL: https://www.facebook.com/business/ads/ad-targeting (cit. a p. 62).

[102] «Meta, azienda social del metaverso». In: (2023). Ultima visita il 26/04/2023. URL: https://about.meta.com/it/ (cit. a p. 60).

[103] Meier Michaela, Vogel Stephan e Grabner Roland. «Going Beyond Intelligence: A Systematic Investigation of Cognitive Abilities and Personality Traits of Experts in Mathematics». In: *Journal of Expertise* 4.1 (mar. 2021), pp. 80–115. URL: https://journalofexpertise.org/articles/volume4_issue1/JoE_4_1_Meier_etal.pdf (cit. alle pp. 23, 24, 57).

[104] Tomas Mikolov et al. «Efficient Estimation of Word Representations in Vector Space». In: (2013). arXiv: 1301.3781 [cs.CL] (cit. a p. 8).

[105] Thomas P. Minka. «Expectation Propagation for approximate Bayesian inference». In: *CoRR* abs/1301.2294 (2013). arXiv: 1301.2294. URL: http://arxiv.org/abs/1301.2294 (cit. a p. 18).

[106] Arnold Mitchell. «The Nine American Lifestyles». In: *American Political Science Review* 78.2 (1984), pp. 515–516. DOI: 10.2307/1963391 (cit. a p. 11).

[107] Y. Moon. «Personalization and personality: Some effects of customizing message style based on consumer personality». In: *Journal of Consumer Psychology* 12.4 (2002), pp. 313–326. URL: 10.1016/S1057-7408(16)30083-3 (cit. a p. 11).

[108] Brenda J. Moscove e Robert G. Fletcher. «Advertising and Marketing in Electronic Commerce». In: (2003). A cura di Hossein Bidgoli, pp. 21–30. DOI: https://doi.org/10.1016/B0-12-227240-4/00002-2. URL: https://www.sciencedirect.com/science/article/pii/B0122272404000022 (cit. a p. 11).

[109] Andreas Mueller. «WordCloud for Python documentation». In: (2020). Ultima visita il 08/04/2023. URL: https://amueller.github.io/word_cloud/ (cit. a p. 39).

[110] Inbal Nahum-Shani et al. «Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support». In: *Annals of Behavioral Medicine* 52.6 (dic. 2017), pp. 446–462. ISSN: 0883-6612. DOI: 10.1007/s12160-016-9830-8. eprint: https:

//academic.oup.com/abm/article-pdf/52/6/446/38625111/abm_52_6_446.pdf. URL: https://doi.org/10.1007/s12160-016-9830-8 (cit. a p. 14).

[111] Nahum-Shani et al. «Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support». In: *Annals of Behavioral Medicine* 52.6 (2018), pp. 446–462. DOI: 10.1007/s12160-016-9830-8 (cit. a p. 12).

[112] Office for National Statistics. «SOC 2010. Previous version of the Standard Occupational Classification». In: (2016). Visitato nei mesi di gennaio e febbraio 2023. URL: https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010 (cit. a p. 17).

[113] H. Nissenbaum. «Privacy in context: Technology, policy, and the integrity of social life». In: (2010) (cit. a p. 14).

[114] Hart R. P. «Verbal style and the presidency: A computer-based analysis». In: (1984) (cit. a p. 6).

[115] T.Y. Tang P. Winoto. «The role of user mood in movie recommendations Expert Systems with Applications». Ver. 37. In: 8 (2010), pp. 6086–6092. URL: 10.1016/j.eswa.2010.02.117 (cit. a p. 12).

[116] Babak Pahlavan. «Google Measurement Partners: Trusted measurement solutions for the entire customer journey». In: (2018). Ultima visita il 27/04/2023. URL: https://blog.google/products/marketingplatform/360/introducing-measurement-partners/ (cit. a p. 64).

[117] Barbara Plank e Dirk Hovy. «Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week». In: (set. 2015), pp. 92–98. DOI: 10.18653/v1/W15-2913. URL: https://aclanthology.org/W15-2913 (cit. alle pp. 10, 13).

[118] Christopher Potts. «Twitter-aware tokenizer». In: (2011). Ultima visita nel mese di gennaio. URL: http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py (cit. a p. 21).

[119] Daniel Preoţiuc-Pietro, Vasileios Lampos e Nikolaos Aletras. «An analysis of the user occupational class through Twitter content». In: (lug. 2015), pp. 1754–1764. DOI: 10.3115/v1/P15-1169. URL: https://aclanthology.org/P15-1169 (cit. alle pp. 17, 18, 27).

[120] Pierce J. R. «Introduction to Information Theory: Symbols, Signals, and Noise, 2nd». In: (1980), pp. 86–87, 238–239 (cit. a p. 6).

[121] Governo del Regno Unito. «Guidance Skilled Worker visa: eligible occupations and codes». In: (2022). Visitato nei mesi di gennaio e febbraio 2023. URL: https://www.gov.uk/government/publications/skilled-worker-visa-eligible-occupations/skilled-worker-visa-eligible-occupations-and-codes (cit. a p. 48).

[122] P. J. Rentfrow. «Statewide differences in personality: Toward a psychological geography of the United States». In: *American Psychologist* 65.6 (2010), pp. 548–558. URL: https://doi.org/10.1037/a0018194 (cit. a p. 11).

[123] Accenture Research. «Cracking the gender code». In: (2016). URL: https://www.accenture.com/_acnmedia/PDF-150/Accenture-Cracking-The-Gender-Code-Report.pdf (cit. alle pp. 1, 79).

[124] Gidi Rubinstein. «The big five among male and female students of different faculties». In: *Personality and Individual Differences* 38.7 (2005), pp. 1495–1503. ISSN: 0191-8869. DOI: https://doi.org/10.1016/j.paid.2004.09.012. URL: https://www.sciencedirect.com/science/article/pii/S0191886904002892 (cit. alle pp. 23, 24, 54).

[125] Berkovsky S. et al. «"Detecting personality traits using eye-tracking data"». In: (2019), pp. 1–12 (cit. alle pp. 10, 46).

[126] Hoppe S. et al. «Eye movements during everyday behavior predict personality traits». In: *Frontiers in Human Neuroscience* 12 (2018), p. 105 (cit. alle pp. 10, 13).

[127] Norsaremah Salleh et al. «The Effects of Neuroticism on Pair Programming: An Empirical Study in the Higher Education Context». In: ESEM '10 (2010). DOI: 10.1145/1852786.1852816. URL: https://doi.org/10.1145/1852786.1852816 (cit. alle pp. 23, 24, 54, 55, 58).

[128] Prema Sampath e Google Stephen Mangan. «Marketing mix models are based in science, but also need a touch of art». In: (2021). Ultima visita il 27/04/2023. URL: https://www.thinkwithgoogle.com/marketing-strategies/data-and-measurement/art-of-marketing-mix-models/ (cit. a p. 64).

[129] Maarten Sap et al. «Developing Age and Gender Predictive Lexica over Social Media». In: (gen. 2014), pp. 1146–1151. DOI: 10.3115/v1/D14-1121 (cit. alle pp. 8, 20, 22, 28, 29, 31, 39).

[130] Stephen M. Schueller, Adrian Aguilera e David C. Mohr. «Ecological momentary interventions for depression and anxiety». In: *Depression and Anxiety* 34.6 (2017), pp. 540–545. DOI: https://doi.org/10.1002/da.22649. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22649. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/da.22649 (cit. a p. 14).

[131] H. Andrew Schwartz et al. «Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach». In: *PLOS ONE* 8 (set. 2013), pp. 1–16. DOI: 10.1371/journal.pone.0073791. URL: https://doi.org/10.1371/journal.pone.0073791 (cit. alle pp. 8, 10, 13, 20–22, 32, 39–44, 47, 53).

[132] Chok Nian Shong. «Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data». In: (2010). URL: http://d-scholarship.pitt.edu/8056/ (cit. a p. 52).

[133] Pushpa Singh et al. «Chapter 5 - Diagnosing of disease using machine learning». In: *Machine Learning and the Internet of Medical Things in Healthcare*. A cura di Krishna Kant Singh et al. Academic Press, 2021, pp. 89–111. ISBN: 978-0-12-821229-5. DOI: https://doi.org/10.1016/B978-0-12-821229-5.00003-3. URL: https://www.sciencedirect.com/science/article/pii/B9780128212295000033 (cit. a p. 30).

[134] Michael R. Solomon. «Consumer Psychology». In: (2004). A cura di Charles D. Spielberger, pp. 483–492. DOI: https://doi.org/10.1016/B0-12-657410-3/00219-1. URL: https://www.sciencedirect.com/science/article/pii/B0126574103002191 (cit. a p. 11).

[135] StatCounter. «Search Engine Market Share Worldwide». In: (2023). Ultima visita il 29/04/2023. URL: https://gs.statcounter.com/search-engine-market-share (cit. a p. 65).

[136] Statista. «Distribution of TikTok users worldwide as of January 2023, by age and gender». In: (2023). Visitato il 05/05/2023. URL: https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution/ (cit. a p. 77).

[137] Statista. «Most popular social networks worldwide as of January 2023, ranked by number of monthly active users». In: (2023). Visitato il 28/04/2023. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (cit. a p. 65).

[138] David A. Swanson. «Applied Demography». In: (2015). A cura di James D. Wright, pp. 839–844. DOI: https://doi.org/10.1016/B978-0-08-097086-8.10513-6. URL: https://www.sciencedirect.com/science/article/pii/B9780080970868105136 (cit. a p. 11).

[139] Costa P. T. e McCrae R. R. «Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. Journal of Personality and Social Psychology». In: 38.4 (1980), pp. 668–678. URL: https://doi.org/10.1037/0022-3514.38.4.668 (cit. a p. 9).

[140] Yla R. Tausczik e James W. Pennebaker. «The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods». In: *Journal of Language and Social Psychology* 29.1 (2010), pp. 24–54. DOI: 10.1177/0261927X09351676. eprint: https://doi.org/10.1177/0261927X09351676. URL: https://doi.org/10.1177/0261927X09351676 (cit. alle pp. 6, 20).

[141] C. Thompson. «The Secret History of Women in Coding». In: *The New York Times Magazine* (2019). URL: https://www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html (cit. a p. 1).

[142] Treccani. «Omnia: definizioni, etimologia e citazioni nel Vocabolario». In: (2023). URL: https://www.treccani.it/vocabolario/ricerca/omnia/ (cit. a p. 45).

[143] Fabrizio Trentacosti. «Come caricare la tua lista clienti su Facebook». In: (2016). URL: https://trucchifacebook.com/facebook/guida/come-caricare-lista-clienti-su-facebook/ (cit. a p. 74).

[144] E.C. Tupes e R.E. Christal. «Recurrent Personality Factors based on Trait Ratings». In: *Journal of personality* 60 (1961), pp. 225–251 (cit. a p. 9).

[145] Yusuke Umegaki e Ayaka Higuchi. «Personality traits and mental health of social networking service users: A cross-sectional exploratory study among Japanese undergraduates». In: *Computers in Human Behavior Reports* 6 (2022), p. 100177. ISSN: 2451-9588. DOI: https://doi.org/10.1016/j.chbr.2022.100177. URL: https://www.sciencedirect.com/science/article/pii/S2451958822000112 (cit. alle pp. 25, 38, 54, 56, 57).

[146] Bakir V. «Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica's Psychographic Profiling and Targeting». In: *Front. Commun.* 5.67 (2020). DOI: doi:10.3389/fcomm.2020.00067. URL: https://www.frontiersin.org/articles/10.3389/fcomm.2020.00067/full (cit. alle pp. 2, 11, 14).

[147] Fernanda B. Viegas et al. «Participatory Visualization with Wordle». In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009). In data corrente (aprile 2023) la piattaforma ufficiale della pubblicazione non è più accessibile, https://www.wordle.net/advanced., pp. 1137–1144. DOI: 10.1109/TVCG.2009.171 (cit. a p. 22).

[148] Pennebaker J. W., Booth R. J. e Francis M. E. «Linguistic inquiry and word count: LIWC [Computer software]». In: (2007) (cit. a p. 6).

[149] Pennebaker J. W. et al. «The development and psychometric properties of LIWC2015». In: (2015) (cit. a p. 6).

[150] Roberts B. W. et al. «What is conscientiousness and how can it be assessed?» In: 50.5 (2014), pp. 1315–1330. URL: https://doi.org/10.1037/a0031109 (cit. alle pp. 25, 37).

[151] F. Wahle et al. «Mobile sensing and support for people with depression: A pilot trial in the wild». In: *JMIR mHealth and uHealth* 4.3 (2016), e111. DOI: 10.2196/mhealth.5960 (cit. a p. 14).

[152] Lucy Lu Wang et al. «Gender trends in computer science authorship». In: *CoRR* abs/1906.07883 (2019). arXiv: 1906.07883. URL: http://arxiv.org/abs/1906.07883 (cit. a p. 1).

[153] Miller LC Wang L. «Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. Health Commun». Ver. 35. In: 12 (2020), pp. 1531–1544. DOI: 10.1080/10410236.2019.1652388 (cit. a p. 12).

[154] Elizabeth Weingarten. «The STEM Paradox: Why Are Muslim-Majority Countries Producing So Many Female Engineers?» In: (2017). URL: https://slate.com/human-interest/2017/11/the-stem-paradox-why-are-muslim-majority-countries-producing-so-many-female-engineers.html (cit. a p. 2).

[155] Michael P. Wilmot e Deniz S. Ones. «A century of research on conscientiousness at work». In: *Proceedings of the National Academy of Sciences* 116.46 (2019), pp. 23004–23010. DOI: 10.1073/pnas.1908430116. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1908430116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1908430116 (cit. alle pp. 25, 26, 37).

[156] Stephan Winter, Ewa Maslowska e Anne L. Vos. «The effects of trait-based personalization in social media advertising». In: *Computers in Human Behavior* 114 (2021), p. 106525. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2020.106525. URL: https://www.sciencedirect.com/science/article/pii/S0747563220302776 (cit. a p. 78).

[157] WordArt.com. «WordArt». In: (2023). Ultima visita il 08/04/2023. URL: https://wordart.com/ (cit. a p. 39).

[158] Amichai-Hamburger Y., Wainapel G. e Fox S. «"On the Internet no one knows I'm an introvert": extroversion, neuroticism, and Internet interaction. Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society». In: 5.2 (2002), pp. 125–128. DOI: DOI:10.1089/109493102753770507. URL: https://pubmed.ncbi.nlm.nih.gov/12025878/ (cit. alle pp. 25, 38, 54, 56, 57).

[159] Liu Y. et al. «RoBerta: A robustly optimized BERT pretraining approach». In: (2019). arXiv: 1907.11692 (cit. a p. 8).

[160] Jun Ye. «Pearson, Spearman, Kendall, Biserial, Tetrachoric and more». In: (2020). URL: https://junye0798.com/post/everythin-you-need-to-know-about-correlation/ (cit. a p. 53).

[161] Weisberg YJ, Deyoung CG e Hirsh JB. «Gender Differences in Personality across the Ten Aspects of the Big Five». In: *Front Psychol* 1.2 (2011), p. 178. DOI: doi:10.3389/fpsyg.2011.00178. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149680/ (cit. alle pp. 25, 57).

[162] M. Younas. «Research challenges of big data». In: *Service Oriented Computing and Applications volume* 13 (2019), pp. 105–107. DOI: https://doi.org/10.1007/s11761-019-00265-x (cit. a p. 12).

[163] YouTube. «Presenta la tua attività a nuovi clienti». In: (2023). Ultima visita il 27/04/2023. URL: https://www.youtube.com/intl/ALL_it/ads/how-it-works/set-up-a-campaign/awareness/ (cit. a p. 64).

[164] YouTube. «Programma di ricerca di YouTube». In: (2023). Ultima visita il 28/04/2023. URL: https://research.youtube/ (cit. a p. 66).

[165] Yong Zheng, Bamshad Mobasher e Robin Burke. «Emotions in Context-Aware Recommender Systems». In: (2016). A cura di Marko Tkalčič et al., pp. 311–326. DOI: 10.1007/978-3-319-31413-6_15. URL: https://doi.org/10.1007/978-3-319-31413-6_15 (cit. a p. 12).

[166] Zygomatic. «Free online Wordcloud generator». In: (2023). Ultima visita il 08/04/2023. URL: https://www.wordclouds.com/ (cit. a p. 39).