

UNIVERSITÀ DEGLI STUDI DI PERUGIA
Dipartimento di Matematica e Informatica

CORSO DI LAUREA MAGISTRALE IN INFORMATICA



Tesi di Laurea

**Verso la Prevenzione del Grooming Online:
Classificazione dell'Arousal Emozionale nelle Prime
Fasi dell'Adescamento**

Laureando:

Marco Amici

Relatori:

Prof. Franzoni Valentina

Dott. Florindi Emanuele

Dott. Polticchia Mattia

Indice

1	Introduzione	1
1.1	Cos'è il grooming?	5
1.2	Problematiche allo stato dell'arte	14
1.3	Obiettivo della Tesi	19
2	Metodologie	21
2.1	Natural Language Processing	21
2.2	Classificazione del linguaggio in classi emotive	35
2.3	Large Language Model	42
2.4	Modello delle Emozioni di Plutchik	49
2.5	Generazione dei Data Set	53
3	Esperimenti	68
3.1	Scelta Modelli e Metriche per la Prossimità Semantica . . .	70
3.2	Scelta modelli e parametri per fine-tuning	73
3.3	Classificazione emozioni primarie e secondarie	76
3.4	Classificazione degli Stati Affettivi Composti	84
3.5	Applicazione al Data set di Grooming	90
4	Risultati e conclusioni	93
4.1	Risultati Embedding e Prossimità Semantica	93
4.2	Risultati del Fine-tuning per le Emozioni Primarie	101
4.3	Risultati Addestramento delle Teste di Classificazione	106
4.3.1	Risultati Classificazione Stati Affettivi Composti . . .	116

4.4 Rilevamento Emozioni nel Grooming	118
4.5 Conclusioni	136
5 Ringraziamenti	141
Bibliografia	143

Elenco delle figure

2.1	Modello delle emozioni di Plutchik	50
2.2	Distribuzione delle etichette in <i>emotion_label2</i> (emozioni primarie).	56
2.3	Distribuzione delle etichette in <i>emotion_label1</i> (emozioni secondarie).	57
2.4	Distribuzione delle etichette in <i>emotion_label1</i> (stati emotivi). .	57
2.5	Similarità semantica emozioni primarie con Sentence T5. . .	59
2.6	Similarità semantica emozioni secondarie con modello Sentence T5.	60
2.7	Similarità semantica stati affettivi composti con modello Sentence T5.	60
2.8	Estratto data set sintetico chat di grooming	67
3.1	Flusso degli esperimenti	68
3.2	Flusso di esecuzione nell’architettura sviluppata.	69
3.3	Flusso di esecuzione dell’esperimento 1.	72
3.4	Diagramma di flusso dell’addestramento e confronto dei vari modelli	75
3.5	Diagramma dettagliato dell’architettura gerarchica	81
3.6	Approccio (1): Flusso con solo controllo sulla soglia di similarità. .	89
3.7	Approccio (2): Flusso con controllo aggiuntivo sulla probabilità delle emozioni primarie predette in Top-3.	90
3.8	Workflow della classificazione finale	92

4.1	Accuracy semantic similarity con CS	96
4.2	Accuracy semantic similarity con FS	97
4.3	Accuracy semantic similarity con RT	97
4.4	Matrice di confusione di Sentence-T5 con $k = 1$ e FS	99
4.5	Matrice di confusione di Sentence-T5 con $k = 3$ e FS	99
4.6	Matrice di confusione di Sentence-T5 con $k = 1$ e CS	100
4.7	Matrice di confusione di Sentence-T5 con $k = 3$ e CS	100
4.8	Matrice di confusione di RoBERTa con LR a 3.00E-05, BS a 16 e WD a 0.001	105
4.9	Matrice di confusione di RoBERTa con LR a 3.00E-05, BS a 16 e WD a 0.01	106
4.10	Matrice di confusione relativa alla testa primaria	107
4.11	Matrice di confusione della testa secondaria per rabbia	108
4.12	Matrice di confusione della testa secondaria per anticipazione	108
4.13	Matrice di confusione della testa secondaria per gioia	109
4.14	Matrice di confusione della testa secondaria per disgusto	109
4.15	Matrice di confusione della testa secondaria per fiducia	110
4.16	Matrice di confusione della testa secondaria per tristezza	110
4.17	Matrice di confusione della testa secondaria per sorpresa	111
4.18	Matrice di confusione della testa secondaria per paura	111
4.19	Matrice di confusione relativa alla testa primaria per le 3 metaclassi	114
4.20	Matrice di confusione della testa secondaria per le 3 metaclassi	114
4.21	Frequenze relative rilevanti per Tags	126
4.22	Frequenze relative rilevanti per Phases	126
4.23	Indici di dominanza rilevanti per Tags	131
4.24	Indici di dominanza rilevanti per Phases	132
4.25	Entropia per Tags	134
4.26	Entropia per Phases	134

Capitolo 1

Introduzione

Il fenomeno del *grooming* (i.e., adescamento) online di minori ha registrato negli ultimi anni un preoccupante aumento, favorito dalla distribuzione di smartphone e console di gioco, con una sempre crescente disponibilità di Internet agli adolescenti più giovani. Le nuove tecnologie e l'anonimato offerto dalle interazioni via chat hanno esposto i giovani a un rischio crescente di manipolazione psicologica con fini di abuso. In questo contesto, le forze dell'ordine e le istituzioni si trovano ad affrontare sfide significative, poiché il fenomeno si sviluppa rapidamente e i predatori online utilizzano strategie che rendono complesso il monitoraggio e il rilevamento delle conversazioni sospette, che è anche parzialmente coperto dalle leggi sulla privacy. Secondo l'ultimo rapporto annuale della Polizia Postale italiana, del 2024 [1], i casi di adescamento online hanno registrato un incremento del 5% rispetto all'anno precedente, con un aumento del 22% delle vittime di età compresa tra i 14 e i 16 anni, dove la fascia maggiormente colpita è quella tra i 10 e i 13 anni, che rappresenta il 55,7% dei casi totali del 2024. I predatori online non agiscono spontaneamente, ma mettono in atto strategie psicologiche mirate per manipolare le vittime. In alcuni casi, utilizzano tecniche di ingegneria sociale, Open Source Intelligence (OSINT) e Social Media Intelligence (SOCMINT) [2] per ottenere informazioni sui minori, guadagnarne la fiducia, abbatterne le difese psicologiche e instaurare relazioni manipolative. La necessità di

accuratezza e sistematicità di tale approccio coinvolge tutti gli stakeholder (i.e., attori) della protezione dei minori, dalle famiglie alle forze dell'ordine, dagli educatori al governo, alle aziende, ai divulgatori.

Uno studio condotto da *Save the Children* nel 2023 [3] ha mostrato che, nel 47,5% dei casi, l'adescatore è un estraneo, nel 23% una persona conosciuta dalla vittima, nel 16,4% un allenatore o una figura autoritaria nella vita del minore, nel 9,8% un educatore o un insegnante, mentre, nel restante 3,3%, un familiare stretto. La vicinanza costante di alcune figure al minore aumenta la difficoltà di rilevare ed affrontare l'adescamento e spesso porta la vittima a non denunciare l'abuso per timore di disgregare gli equilibri della propria rete sociale di supporto.

Una ricerca del 2021 [4] ha mostrato che quasi la metà dei minori che avevano condiviso immagini intime, successivamente rese pubbliche, lo aveva fatto con qualcuno conosciuto solo online, mentre il 40% con una persona ritenuta erroneamente un adulto. Questo dato evidenzia come l'adescamento online non si limiti più al contatto fisico, ma sfrutta la comunicazione anonima. Un ulteriore studio del 2020 [5] aveva già rivelato che, su un campione di 1133 partecipanti in età minorile al sondaggio, un quarto aveva conversato con adulti sconosciuti online e il 65% di questi aveva ricevuto sollecitazioni sessuali. Inoltre, il 23% ha ricordato di aver avuto una lunga conversazione intima in una chat online con un adulto, secondo gli schemi tipici di adescamento sessuale.

La limitata consapevolezza digitale, soprattutto tra i genitori o chi ne fa le veci, rappresenta un fattore di rischio critico. Una recente intervista, condotta in uno studio del 2021 [6] su un campione di 19 genitori con figli tra i 13 e i 17 anni, ha evidenziato che oltre il 60% degli intervistati è consapevole del fenomeno del grooming online, ma evita di parlarne con i figli a causa della sensibilità dell'argomento, anziché rendersi disponibile alla prevenzione.

Se diverse normative, come il Regolamento Generale europeo sulla Protezione dei Dati (GDPR) [7] o il regolamento sulla privacy di Hong Kong [8], rappresentano presidi essenziali per la tutela del diritto alla privacy, tuttavia alcune di esse presentano dei limiti per la prevenzione del grooming online. Ad esempio, il GDPR impone vincoli stringenti sulla raccolta ed elaborazione di dati sensibili relativi ai minori, in assenza del consenso esplicito dei tutori legali: tali restrizioni, sebbene giustificate dal principio di protezione dei soggetti vulnerabili, possono ostacolare le attività di monitoraggio precoce e compromettere l'efficacia degli interventi preventivi, ritardando l'identificazione tempestiva di situazioni potenzialmente riconducibili ad abusi.

Studi recenti su prevenzione e controllo [9] evidenziano invece la rilevanza cruciale dell'intercettazione precoce di situazioni potenzialmente a rischio, quale strumento efficace per la prevenzione di abusi. Tuttavia, tali approcci si scontrano con significative limitazioni anche di natura tecnica, tra cui l'adozione sempre più diffusa della crittografia end-to-end nei servizi di messaggistica, che, pur garantendo la riservatezza delle comunicazioni, non permettendo l'accesso di terze parti ai contenuti, riduce la possibilità di attuare forme di monitoraggio proattivo e controllo, sia da parte delle piattaforme che -in parte- delle stesse forze dell'ordine.

L'adozione di tecnologie avanzate come l'intelligenza artificiale potrebbe certamente rappresentare un valido strumento per individuare tempestivamente possibili segnali emotivi e linguistici riconducibili a schemi tipici di grooming, nonché per simulare l'adescamento a scopo educativo e di sensibilizzazione. L'approccio proposto in questa tesi di ricerca si basa sull'analisi delle interazioni digitali al fine di sviluppare un sistema capace di rilevare in tempo reale comportamenti sospetti, nel rispetto delle normative vigenti. Questo sistema si avvale dell'identificazione di specifiche caratteristiche linguistiche dei messaggi, permettendo di individuare attività potenzialmente manipolatorie prima che si trasformino in abuso. In particolare, l'analisi

delle dinamiche psicologiche del grooming, soprattutto nelle sue fasi iniziali, è cruciale per prevenire le interazioni traumatizzanti.

In questo contesto, vengono qui esplorati modelli e tecniche legati al Natural Language Processing (NLP), attraverso l'uso di Large Language Models (LLM) e strumenti per l'analisi semantica del testo. L'obiettivo del sistema studiato in questa tesi di ricerca è duplice: da un lato, fornire strumento utili a monitoraggio e segnalazione di interazioni sospette; dall'altro, fornire uno strumento formativo per migliorare consapevolezza, sensibilità e capacità di intervento, sia degli adulti che dei minori, nella loro protezione. Tale sistema di rilevamento si propone, infatti, come un supporto per la sensibilizzazione e per l'identificazione precoce di situazioni potenzialmente a rischio, integrato in un approccio multidisciplinare tra informatica, neuroscienze e aspetti etico-legali. Si ritiene, infatti, che solo attraverso la sinergia tra strumenti tecnologici, interventi educativi, supporto psicologico e collaborazione tra istituzioni, sia possibile costruire un contesto realmente efficace per la prevenzione e la tutela dei minori in ambienti digitali.

La nostra tesi è articolata in quattro capitoli: dopo un'analisi dettagliata, nel seguito di questo capitolo, del fenomeno del grooming e delle problematiche allo stato dell'arte, verranno presentate nel capitolo 2 le metodologie fondamentali per il raggiungimento dell'obiettivo finale, tra cui l'architettura Transformer usata nei Large Language Model; le tecniche di fine-tuning di modelli generativi, che ci serviranno per la generazione di un data set e il modello delle emozioni di Plutchick; nei capitoli 3 e 4 verranno poi presentati e discussi gli esperimenti e i relativi risultati, partendo dalla generazione del data set e proseguendo con il confronto e la scelta dei modelli per l'analisi della prossimità semantica e per la classificazione gerarchica delle emozioni da associare alle fasi del grooming, passando per l'implementazione del modello finale di classificazione, addestrato sul data set ad hoc, generato sinteticamente, poi applicato al data set annotato con le fasi e con tag sulle azioni specifiche del grooming.

1.1 Cos'è il grooming?

Il grooming è definito come un processo graduale e intenzionale in cui un adulto instaura una relazione di fiducia con un minore, anche attraverso mezzi digitali, con lo scopo di manipolare, sfruttare o abusare sessualmente della vittima [10]. Questo processo si sviluppa attraverso fasi progressive. Il groomer adatta strategicamente il proprio approccio in base all'età, al contesto socio-culturale e alle fragilità individuali del minore, presentandosi inizialmente come figura comprensiva e affabile, per poi far evolvere la relazione verso dinamiche sempre più manipolatorie, fino al ricatto.

Il termine "grooming" deriva dall'inglese "to groom", che significa "prendersi cura": in origine usato in etologia, per riferirsi alla cura di gruppi di animali, venne in seguito adottato in psicologia e criminologia per descrivere il processo di "preparazione" della vittima da parte dell'abusante.

La peculiarità di questo fenomeno risiede nella sua natura subdola e multebole: l'isolamento sistematico della vittima dai suoi contesti relazionali primari (famiglia, amici, contesti educativi come scuola e gruppi sportivi) e la graduale normalizzazione di comportamenti inappropriati compromettono la capacità del minore di riconoscere la situazione di abuso. Sebbene tradizionalmente associato alla sfera sessuale, l'abuso può manifestarsi attraverso diverse forme di sfruttamento, tra cui la manipolazione emotiva, l'induzione all'autolesionismo e il coinvolgimento in attività illecite. Le conseguenze psicologiche sulle vittime sono particolarmente gravi e durature, partendo da disturbi d'ansia, depressione, disturbo post traumatico da stress e alterazioni nella capacità di stabilire relazioni sane che -se non trattate- resteranno, permanenti, in tutta la vita della vittima, fino ad arrivare a loop critici di colpa-vergogna che portano al suicidio.

Il Grooming Online

Nell'era digitale, il grooming online ha assunto caratteristiche sempre più sofisticate, rendendo il fenomeno particolarmente insidioso e complicato da gestire: l'ampia diffusione delle piattaforme virtuali, quali social network, giochi online e applicazioni di messaggistica, frequentate da un'utenza sempre più giovane, ha ampliato le opportunità di azione per i predatori, i quali possono facilmente eludere le tradizionali barriere fisiche e sociali. Avvalendosi dell'anonimato, con la possibilità di creare identità fittizie, i groomer instaurano dinamiche manipolative che complicano la propria individuazione; l'evoluzione tecnologica, inoltre, ha introdotto ulteriori elementi di criticità: possiamo plausibilmente aspettarci che anche i modelli di intelligenza artificiale generativa possano divenire un supporto ai groomer.

Dal punto di vista giuridico, la risposta legislativa, sebbene presente (in Italia con l'art. 609-undecies c.p.) [11], fatica a tenere il passo con la costante evoluzione e mutazione delle tecnologie usate per l'adescamento. La particolare difficoltà nell'identificazione dei predatori digitali, dovuta all'uso di strumenti di anonimizzazione e alla natura transnazionale del Web, rappresenta oggi una delle sfide più significative.

I rischi associati al grooming online sono molteplici: oltre allo sfruttamento sessuale diretto, le vittime possono subire ulteriori danni anche in fasi intermedie, come la diffusione non consensuale di materiale intimo (e.g., revenge porn), la manipolazione emotiva e, nei casi più estremi, il rapimento al fine di sfruttamento o traffico di minori.

Le Fasi del Grooming

Il processo di grooming si sviluppa attraverso diverse fasi. In questo lavoro di tesi, si è adottata la categorizzazione in sei fasi e quattro categorie operative, ciascuna caratterizzata da specifiche strategie manipolative adottate dal predatore per ottenere il controllo della vittima. Sebbene interconnesse, le

fasi del grooming si distinguono per le azioni concrete messe in atto e per l'approccio utilizzato nel consolidare l'influenza sulla vittima.

La psicologa Rachel O'Connell individua le sei fasi fondamentali, che descrivono il progressivo coinvolgimento emotivo e psicologico del minore [12] e le analizza, evidenziando le dinamiche con cui il predatore costruisce un rapporto di fiducia per perseguire i propri scopi illeciti:

1. Contatto (Friendship Forming)

In questa fase iniziale, il predatore stabilisce il primo contatto con la vittima, ad esempio attraverso piattaforme digitali quali social network, applicazioni di messaggistica o giochi online. L'approccio è studiato per apparire innocuo e spontaneo, spesso limitandosi inizialmente ad un saluto o a una richiesta di amicizia, in modo da non destare sospetti. Il predatore si presenta poi come una figura amichevole, empatica e disponibile, talvolta fingendosi un coetaneo o una persona in cerca di compagnia. Per consolidare la fiducia, il groomer fingerà di avere interessi comuni con la vittima, riferendosi a tematiche affini, come gusti musicali, esperienze scolastiche o passioni condivise. Parallelamente, avvierà una raccolta progressiva di informazioni personali, indagando su età, scuola e attività extrascolastiche, abitudini e preferenze del minore. Il groomer si potrà anche spingere a richiedere foto della vittima, con il duplice scopo di valutarne l'aspetto fisico e di identificarla in altri canali o nella vita reale.

2. Costruzione della relazione (Relationship-Forming)

In questa fase, il predatore consolida gradualmente la relazione con la vittima attraverso un'intensa comunicazione, ponendosi come figura empatica e affabile. Adottando strategie di manipolazione emotiva, il criminale incoraggerà dunque la vittima a confidare aspetti personali della propria vita, come problematiche familiari, sentimenti di solitudine o difficoltà relazionali, dimostrando apparente sintonia e

sostegno e rafforzando le convinzioni negative che rappresentano i bisogni a cui possa fornire una risposta affettiva. Attraverso tecniche di validazione emotiva e isolamento indiretto, il predatore aumenta gradatamente così la dipendenza psicologica della vittima, facendo leva sulle sue insicurezze per rafforzare il loro legame. Progressivamente, la relazione viene percepita dalla vittima come unica e speciale, con il predatore che assume il ruolo di confidente privilegiato, sostituendosi alle figure di riferimento tradizionali, in particolare a famiglia e amici. Questo processo, ben documentato in letteratura [13], genera un'arousal emotivo sempre più intenso, uno stato di continua attivazione psicofisiologica che coinvolge componenti emotive e neurobiologiche, in particolare mediante l'attivazione del sistema limbico, implicato nella regolazione affettiva, nella memoria emotiva e nei meccanismi di attaccamento. Il legame affettivo con il manipolatore viene progressivamente interiorizzato, finché il groomer diventa la figura di riferimento privilegiata dalla vittima. L'arousal emotivo prolungato compromette infatti le capacità cognitive della vittima, inibendo il pensiero critico e favorendo l'instaurarsi della dipendenza affettiva, che a sua volta aumenta la vulnerabilità della vittima, rendendola più incline ad accettare e normalizzare comportamenti manipolativi. La vittima è ormai agganciata.

3. Valutazione del rischio (Risk Assessment)

Una volta instaurato il rapporto di fiducia, il predatore avvia una sistematica valutazione del rischio di esposizione, indagando con modalità apparentemente innocue il contesto familiare e sociale della vittima. Attraverso domande indirette e conversazioni mirate, raccoglie informazioni cruciali sulla struttura familiare, sulle abitudini digitali e sul grado di sorveglianza esercitato dagli adulti. Il predatore adotterà strategie manipolative volte a inculcare nella vittima la necessità di mantenere la relazione segreta, enfatizzando il carattere esclusivo del

legame e sminuendo l'importanza delle figure di riferimento. Tale processo, definito generalmente come "strategia di occultamento relazionale" [14], rappresenta una fase critica nel modus operandi del predatore, poiché consente di identificare le vittime più vulnerabili e, al contempo, predisporre le condizioni per un'eventuale escalation dell'abuso. In questa fase, una presenza attenta da parte della famiglia può ancora agire da protezione per il minore e da deterrente per il groomer, evitando il danno.

4. Esclusività del rapporto (Exclusivity)

Una volta accertato che la relazione può proseguire senza interferenze esterne, il predatore intensifica il senso di esclusività e segretezza. In questa fase, la manipolazione emotiva diventa particolarmente evidente: il predatore fa leva su strategie persuasive volte a far sentire la vittima ascoltata e apprezzata, ma anche a rendere la propria presenza necessaria. L'obiettivo principale è l'ulteriore costruzione di un rapporto di esclusività del proprio rapporto con la vittima, isolandola da altri legami ed inducendola a percepire la difficoltà a poter dare fiducia ad altre persone. Attraverso questo processo, il predatore consolida il controllo psicologico sulla vittima, sfruttando sentimenti di fiducia, affetto, necessità e segretezza, per ottenere maggiore potere sulla relazione e maggior libertà di agire indisturbato.

5. Desensibilizzazione e sessualizzazione (Sexualization)

In questa fase, il predatore introduce un processo di graduale escalation dei contenuti sessuali, per passare da richieste apparentemente innocue ad altre più esplicite, come descrizioni di fantasie sessuali, facendo poi pressione per ottenere materiale sempre più esplicito della sfera intima della vittima, passando per fasi di progressiva normalizzazione degli atti eventualmente percepiti come disturbanti.

Questo tipo di manipolazione è simile a quello perpetrato da figure

narcisiste anche nelle relazioni con adulti: con un mix di seduzione e coercizione psicologica, il predatore sfrutta il legame emotivo precedentemente costruito per manipolare la vittima, giocando sulla sua vulnerabilità e sulla percezione di una relazione speciale che ormai le è necessaria. Man mano che la vittima si trova sempre più coinvolta, il predatore intensifica le richieste, passando a vere e proprie estorsioni di materiale esplicito e, se necessario, minacciando di diffondere eventuali immagini già ottenute. Questa tecnica, definita come "sexortion" [15], non solo consente al predatore di ottenere ulteriore materiale compromettente, ma rafforza il controllo psicologico sulla vittima, aumentandone la dipendenza dalla relazione e riducendo la possibilità di denuncia, in un circolo vizioso di manipolazione e abuso che può protrarsi nel tempo. È in questa fase che la vittima subisce gli atti abusanti che le lasceranno segni profondi e duraturi.

6. Coercizione e controllo (Coercion and Control)

Nella fase conclusiva del processo di grooming, il predatore esercita il suo dominio sulla vittima attraverso un sistematico ricatto psicologico ed emotivo. Avvalendosi del materiale compromettente ottenuto nelle fasi precedenti (e.g., immagini intime, confessioni personali), il predatore intensifica le minacce di diffusione, strutturando un vero e proprio sistema di coercizione fondato sulla paura e sul ciclo critico di vergogna e senso di colpa. La vittima, ormai intrappolata, sviluppa la paralisi decisionale che sempre accompagna un intenso arousal emotivo, impedita sia nell'interrompere la relazione, sia nel denunciare gli abusi. Il predatore, dal canto suo, aumenta le richieste di materiale esplicito o incontri fisici, sfruttando la vulnerabilità emotiva della vittima e la sua percezione di isolamento da cui deriva la necessità del rapporto. Questa fase rappresenta l'apice del processo manipolatorio, in cui il potere del predatore si cristallizza. In questa fase, la vittima non riesce più ad uscire dall'abuso senza chiedere aiuto, fino al burnout o

a conseguenze più estreme.

Secondo il CyberTipline Report 2022 del National Center for Missing & Exploited Children (NCMEC) [16], inoltre, meno dell'1% dei casi di sfruttamento sessuale online è stato segnalato direttamente dal pubblico (comprese le vittime), mentre il restante 99% delle segnalazioni proviene dai fornitori di servizi elettronici. Questo dato sottolinea da un lato l'efficacia distruttiva delle strategie di dominazione predatoria, dall'altro l'importanza e l'efficacia delle tecniche di prevenzione e protezione che possono essere integrate nelle piattaforme digitali.

Nel nostro lavoro, le sei fasi del processo di grooming online, così come appena descritte, sono ricondotte analiticamente al modello più sintetico composto da quattro categorie operative:

1. Targeting and Gaining Trust (Individuazione e Acquisizione della Fiducia),
2. Fulfilling Needs (Soddisfacimento dei Bisogni),
3. Isolation (Isolamento),
4. Sexualizing Relationship (Sessualizzazione della Relazione).

Questa riorganizzazione concettuale mantiene la struttura dl fenomeno, senza perdita di rappresentazione della complessità psicologica e relazionale del processo di manipolazione. Allo stesso tempo, si rivela congruente con le fonti esistenti di dati annotati sul grooming, da cui potremo generare un nostro data set.

I fattori determinanti nel Grooming Online

Il fenomeno del grooming online è strettamente legato alla vulnerabilità dei minori, una condizione influenzata da molteplici fattori, sia interni che esterni. Per comprendere meglio le cause di questa esposizione, è essenziale analizzare i principali fattori di rischio che aumentano la predisposizione dei minori alla

manipolazione online. Classifichiamo tali fattori in tre categorie principali, ciascuna delle quali incide significativamente sul livello di vulnerabilità del minore nei confronti degli adescatori:

1. Fattori individuali

Legati alle caratteristiche psicologiche ed emotive del minore. Il minore potrà cercare di colmare eventuali lacune emotive o relazionali attraverso l'interazione con adulti, trovando quindi anche nell'adescatore una figura che sembra in grado di offrire comprensione, affetto e sostegno.

2. Fattori familiari

Un contesto familiare instabile, caratterizzato da conflitti interni, carenza di supporto affettivo o insufficiente attenzione, può indurre i minori a cercare altrove la sicurezza che non trovano in ambito familiare. Ugualmente critiche divengono quindi la scarsa comunicazione tra genitori e figli e la mancanza di supervisione delle attività online facilitano la formazione di legami virtuali con estranei, dove l'assenza di una corretta educazione digitale e la scarsa consapevolezza dei pericoli online contribuiscono ad aumentare il rischio di un'esposizione inconsapevole agli approcci manipolativi degli adescatori.

3. Fattori sociali

Legati al contesto in cui il minore cresce. Un ambiente sociale privo di reti di supporto e una bassa integrazione tra pari (i.e., per il minore, le amicizie) possono accentuare la solitudine, spingendo il minore a cercare connessioni virtuali. La carenza di relazioni significative può rendere il giovane vulnerabile all'adescamento da parte di predatori digitali. Allo stesso modo, anche in ambienti sociali maggiormente attivi, laddove si creassero zone temporali in cui il minore è lasciato solo (e.g., situazioni in cui in famiglie abbienti il minore si trova da solo in momenti di noia e cerca di contrastarla con conversazioni e giochi online senza supervisione), si lascia spazio ad occasioni di grooming.

Un ulteriore aspetto di fondamentale rilevanza è rappresentato dai fattori protettivi che contribuiscono a ridurre la vulnerabilità dei minori nei confronti del fenomeno del grooming online, opponendosi ai fattori di rischio esaminati. Tra questi, la presenza di una rete di supporto sociale, costituita da relazioni affettive significative con familiari, coetanei e figure adulte di riferimento, riveste un ruolo centrale. Inoltre, la personale resilienza, intesa come la capacità dell'individuo di affrontare situazioni avverse e di sviluppare risposte emotive adeguate e funzionali, rappresenta un ulteriore elemento protettivo di rilievo, in grado di rafforzare la capacità dei minori di riconoscere e fronteggiare situazioni potenzialmente pericolose.

La vulnerabilità al grooming online dipende, quindi, dalla combinazione di fattori di rischio e fattori protettivi e rafforzare questi ultimi può ridurre significativamente le possibilità che un minore sia vittima di grooming. A volte, poche semplici indicazioni, insieme ad un rapporto di fiducia con un adulto nella rete di cura, sono sufficienti a scongiurare l'efficacia del primo aggancio.

Manipolazione Emotiva nei messaggi di testo

Il grooming online si configura come un processo manipolativo articolato, in cui il predatore costruisce la sua relazione prevalentemente con messaggi di testo. Lessico rassicurante, simulazione di intimità e strategie discorsive mirano a normalizzare l'influenza del predatore, mascherando intenti coercitivi sotto una patina di positività. Tecniche di manipolazione emotiva come il gaslighting [17] o l'isolamento emotivo vengono implementate attraverso precise scelte lessicali e narrative che distorcono la percezione della realtà. L'analisi di emozioni e stati affettivi dal testo rivela qui tutta la sua utilità, permettendo di individuare pattern emotivi che segnalano la distorsione cognitiva indotta. Una categorizzazione efficace delle strategie manipolative

distingue tra:

- **Tecniche dirette:** manifestazioni esplicite di affetto o preoccupazione che mirano a ottenere compliance immediata, spesso attraverso iperboli emotive.
- **Tecniche indirette:** l'adescatore non esplicita il proprio interesse, ma gioca sulle emozioni della vittima in modo più sottile, facendo leva su sentimenti di carenza o desiderio non esplicitamente dichiarato, facendo crescere nella vittima un bisogno emotivo per il quale il predatore si pone, poi, come unica soluzione.

La sinergia tra queste modalità crea il cosiddetto *doppio legame* [18] e la vittima viene alternativamente gratificata e destabilizzata, in un ciclo che aumenta la dipendenza.

1.2 Problematiche allo stato dell'arte

La progettazione di soluzioni efficaci per il rilevamento del grooming online presenta sfide complesse. La ricerca e lo sviluppo di nuove soluzioni dovranno essere aggiornate alle tecniche e agli strumenti in costante evoluzione a disposizione dei groomer. Nel nostro lavoro di ricerca, alcune difficoltà sono rimaste centrali nel nostro focus.

Sfide nel data set

Il problema principale riguarda la difficoltà di reperire conversazioni autentiche di grooming: le normative sulla privacy e il carattere sensibile di queste interazioni rendono quasi impossibile accedere a dialoghi reali tra predatori e vittime. Di fatto, l'unica risorsa attualmente disponibile allo stato dell'arte sembra essere il dataset denominato *dataset_sintetico_grooming*, realizzato da Ludovico Guercio [19]. In questo dataset, ogni conversazione è annotata con uno specifico tag, appartenente a una delle seguenti categorie: *activities*,

personal information, compliment, relationship, reframing, communicative desensitization, isolation e approach.

La natura stessa del grooming online porta inoltre ad un linguaggio adattato al canale, al profilo simulato dal groomer ed a quello reale della vittima, il tutto con le variazioni relative alle diverse fasi del grooming (v. sezione 1.1). Il rilevamento di potenziali casi di adescamento, in questo contesto di variabilità, implica la necessità di usare modelli in grado di gestire dati di diversa natura e di cogliere pattern complessi. Le soluzioni più promettenti combinano tecniche avanzate di linguistica computazionale con modelli in grado di analizzare il contesto delle conversazioni. Nel nostro lavoro, si è deciso di usare architetture basate sull'attention mechanism [20] per identificare gli schemi comunicativi del grooming. Resta fondamentale sviluppare metodi di validazione rigorosi, con un confronto sistematico tra gli strumenti disponibili e la valutazione dei più adatti nello scenario d'uso.

Sul piano legale, in Italia l'adescamento è considerato reato fin dalle prime fasi. Tuttavia, la difficoltà risiede nel fatto che le prime fasi del grooming a volte non contengono contenuti esplicativi tali da identificare subito gli illeciti, il che ostacola un intervento tempestivo, creando un gap tra l'individuazione di comportamenti sospetti e la possibilità di azione legale. Questo permette alla minaccia di evolversi prima che possano essere adottate misure adeguate. A questo riguardo, il nostro lavoro si concentra sulla fase di isolamento, dove l'arousal emotivo è tale da ipotizzare di poter identificare il grooming prima che l'atto dell'abuso traumatizzante abbia luogo, ponendo in atto dunque una strategia preventiva.

Stato dell'arte dell'analisi testuale

Le strategie di base per il rilevamento del contenuto testuale sono basate sull'analisi lessicale, attraverso dizionari di parole chiave, termini frequentemente associati a comportamenti specifici come quelli predatori [21]. Questo

approccio, pur rappresentando un tentativo strutturato ed efficace di automazione, risulta tuttavia eccessivamente rigido, limitato a uno specifico linguaggio e lento da aggiornare.

Un secondo elemento, cruciale nello sviluppo di sistemi di rilevamento automatici, è rappresentato dal preprocessing linguistico, ovvero dall’insieme di trasformazioni che rendono il testo analizzabile con tecniche automatizzate. I metodi standard di analisi testuale, come il Term Frequency-Inverse Document Frequency (TF-IDF) [22], si sono dimostrati efficaci nell’individuare l’importanza relativa dei termini in un documento, anche senza far riferimento a dizionari. Tuttavia, il contesto del singolo documento è spesso insufficiente per l’analisi semantica.

Modelli tradizionali di Machine Learning come le Support Vector Machine (SVM) e la Logistic Regression (LR) [23] sono impiegati per la classificazione semantica di messaggi, ma la loro efficacia è limitata dalla difficoltà nel rappresentare la dimensione temporale e narrativa delle conversazioni.

Per superare le limitazioni di queste strategie nell’analisi semantica e contestuale, sono stati introdotti sistemi specializzati nella similarità semantica e modelli di rappresentazione distribuita come Word2Vec [24] e GloVe [25] o le loro evoluzioni [26], che proiettano le parole in spazi vettoriali continui preservando le loro proprietà semantiche. Tuttavia, anche questi approcci si sono rivelati parziali nel cogliere il significato dinamico e contestuale del linguaggio, soprattutto in ambienti conversazionali complessi. I modelli di similarità semantica basati sul Web [27] presentano al contrario un’alta flessibilità della semantica, che evolve con l’evoluzione del Web, con il vantaggio di essere sempre aggiornati, ma la variabilità dei risultati non permette una validazione stabile, poiché l’output dipende principalmente dall’aggiornamento in tempo reale del Web, ma in seconda battuta anche dal motore di ricerca usato per l’analisi delle frequenze, nonché dalle misure di prossimità specifiche dei diversi approcci alla semantica in questo ambito.

Nel campo del deep learning, modelli ricorrenti come le Long Short-Term Memory (LSTM) [28] sono capaci di mantenere informazioni nel tempo e di catturare relazioni sequenziali tra le frasi. Parallelamente, le Convolutional Neural Networks (CNN), adattate all’elaborazione del testo, permettono l’identificazione di pattern rilevanti sia sintattici che semantici [29]. Anche la struttura morfo-fraseologica del testo scritto [30] fa la sua parte per la comprensione dei pattern linguistici.

Un avanzamento è certamente l’architettura Transformer, recentemente introdotta da Vaswani et al., nel 2017 [31] e, in particolare, del modello Bidirectional Encoder Representations from Transformers (BERT) [32], capace di considerare il contesto bidirezionale delle parole in una frase. I modelli basati sui Transformer hanno poi raggiunto prestazioni eccellenti nella comprensione delle intenzioni comunicative, tramite i Large Language Model (LLM) [33]. Tuttavia, la qualità dei risultati dipende in maniera sostanziale dalla disponibilità di informazioni rappresentative nei dati di training, con la possibilità di cadere in errori grossolani (i.e., *allucinazioni*) e di ottenere risposte troppo variabili in base al parametro di casualità della generazione di risposte (i.e., *temperature*).

Negli studi più recenti, si stanno affermando approcci ibridi che integrano deep learning, analisi semantica e metodi contrastivi. Le tecniche di contrastive learning [34], che apprendono a discriminare tra esempi simili e dissimili, si sono rivelate promettenti nella rilevazione di comportamenti più sofisticati. Tecniche di clustering comportamentale e network analysis [35] risultano essenziali per rilevare attività sospette non immediatamente visibili attraverso l’analisi linguistica. L’integrazione di tali informazioni all’interno di framework di Knowledge Discovery (KD), insieme a strategie più basilari come l’uso di metadati contestuali (e.g., timestamp dei messaggi, durata delle conversazioni, ritmo di digitazione) potrebbe offrire una comprensione più profonda e sistematica delle dinamiche comportamentali.

Per l’analisi del grooming, una proposta significativa è rappresentata dallo studio di Chehbouni et al. pubblicato nell’articolo "Enhancing Privacy in the Early Detection of Sexual Predators Through Federated Learning and Differential Privacy" [36], che introduce una pipeline basata su federated learning e privacy differenziale. L’approccio proposto da Chehbouni et al. consente il rilevamento precoce di comportamenti predatori senza la necessità di trasferire dati sensibili (in particolare di minorenni) verso server centralizzati, riducendo così il rischio di violazioni della privacy. Attraverso l’uso combinato di Federated Learning e Differential Privacy, è possibile preservare la riservatezza dei dati mantenendo al contempo buoni livelli di accuratezza predittiva, con una marginale perdita di utilità, rendendo tale metodologia promettente per applicazioni reali su larga scala. Tuttavia, il modello affronta il problema in termini esclusivamente binari, classificando ogni segmento conversazionale come *grooming* o *non grooming*, basandosi su un dataset esistente (PANC dataset) costituito da esempi annotati con etichette binarie. L’estrazione delle rappresentazioni testuali avviene mediante l’encoding [CLS] di BERT pre-addestrato, senza effettuare fine-tuning specifico per il task, mentre la classificazione viene eseguita tramite un modello supervisionato di tipo logistic regression, operando su rappresentazioni semantiche generali senza modellare in modo esplicito le dinamiche psicologiche o emotive della conversazione. Il nostro lavoro si differenzia introducendo un approccio fine-grained, orientato alla caratterizzazione psicologica e comunicativa dell’interazione. In particolare, per la rilevazione di comportamenti predatori, viene proposta l’identificazione delle emozioni veicolate dai pattern manipolativi nel processo di adescamento, basandosi su dataset generati sinteticamente.

Un altro interessante studio è quello dell’articolo "Helpful or Harmful? Exploring the Efficacy of Large Language Models for Online Grooming Prevention" [37], che analizza l’uso di modelli linguistici generativi (LLM) in oltre 6.000 interazioni simulate, con l’obiettivo di valutare la loro capacità di

fornire consigli preventivi e riconoscere scenari di grooming in conversazioni tra predatori e "decoy children" (adulti che si fingono minori). I risultati rivelano limiti sostanziali nella coerenza e affidabilità delle risposte prodotte da LLM, che in alcuni casi generano contenuti potenzialmente dannosi o fuorvianti, anziché guidare il minore verso una protezione. Da questo studio desumiamo che un sistema di rilevamento debba essere pensato per alzare delle red flag, dove sarà poi l'utente umano o lo stakeholder di riferimento a valutare caso per caso la strategia di intervento nel caso reale.

Infine, lo studio "Evaluating Language Models on Grooming Risk Estimation Using Fuzzy Theory" [38] di Bihani, Ringenbergs e Rayz, propone un approccio ibrido che integra Sentence-BERT (SBERT) con la logica fuzzy per attribuire un punteggio sfumato di rischio alle conversazioni sospette. Questo metodo si distingue per la capacità di rappresentare in modo più flessibile il grado di pericolosità dei messaggi, anziché adottare classificazioni rigorose. Tuttavia, gli autori evidenziano difficoltà persistenti nel riconoscere il linguaggio implicito o non sessualmente esplicito, sottolineando così i limiti di un approccio di classificazione diretta.

Questi recenti studi rappresentano avanzamenti rilevanti nell'applicazione dell'intelligenza artificiale al contrasto del grooming online, pur evidenziandone i limiti, sia algoritmici che etici (e.g., legati alla privacy e dall'affidabilità dei modelli generativi).

1.3 Obiettivo della Tesi

Lo stato dell'arte evidenzia come research gap i limiti algoritmici degli approcci già presenti e la difficoltà a reperire dati sul grooming.

La nostra proposta di ricerca è quella di usare l'Affective Computing per il rilevamento dell'arousal emozionale nelle fasi critiche del grooming. Data però la complessità delle dinamiche emozionali del grooming, non riteniamo promettenti i modelli emozionali di base come quello di Ekman [39]. D'altra

parte, anche data set basati su modelli più complessi come quello di Plutchick [40], potenzialmente promettente, sono scarsi e poco rappresentativi.

Per una adeguata knowledge base per l’addestramento del nostro sistema, applichiamo la stessa tecnica che è stata usata per generare ad hoc dati di grooming nello studio di Ludovico Guercio [19], usando dati esistenti come documentazione per applicare la tecnica della Retrieval-Augmented Generation (RAG) [41] alla generazione di un nuovo data set per le emozioni tramite LLM.

L’obiettivo di questa tesi di ricerca è dunque sviluppare un sistema avanzato di supporto al rilevamento del grooming online, basato su tecniche di affective computing e sull’impiego di modelli di deep learning per la classificazione e di Large Language Models (LLM) per la generazione di dati di addestramento. Il sistema è pensato per l’applicazione in simulatori utili in campo educativo e di sensibilizzazione. Inoltre, lo stesso sistema potrebbe essere integrato -a seguito di apposite convenzioni- nelle piattaforme di chat, per monitorare le conversazioni digitali al fine di identificare tempestivamente comportamenti sospetti riconducibili a strategie di adescamento e segnalarle prima che possano evolvere in forme conclamate di abuso.

L’analisi si concentrerà dunque principalmente sulla dimensione emotiva delle interazioni testuali. Il sistema si focalizzerà in particolare sul riconoscimento progressivo dei comportamenti predatori, con un’attenzione specifica alle dinamiche emozionali evocate nelle diverse fasi del grooming: in tale ricerca verranno studiati i pattern emozionali, per verificare se esistono effettive correlazioni con le fasi del grooming. Un secondo contributo significativo è il data set sintetico etichettato tramite tecniche di in-context learning, che vanno ad integrare gli scarsi dati disponibili, sui quali è stato generato.

Capitolo 2

Metodologie

In questo capitolo parliamo delle principali metodologie usate nel nostro lavoro di analisi del linguaggio delle conversazioni digitali per l'identificazione precoce del grooming (v. sez. 1). Saranno esaminati, in primo luogo, i fondamenti teorici e computazionali che sono alla base della rappresentazione e della comprensione del linguaggio naturale da parte delle macchine, con un focus sui modelli distribuzionali, semantici e neurali [42] più rilevanti. A tal fine, ci concentriamo su tecniche avanzate di modellazione semantica per l'analisi del contenuto testuale (e.g., word e sentence embedding [43]) e strumenti per la classificazione delle emozioni, basati sul modello di Plutchik [40]. Particolare attenzione viene rivolta all'identificazione di pattern linguistici ed emotivi che possano fungere da indicatori predittivi di condotte manipolative. L'approccio adottato integra anche componenti di similarità semantica [44], per la rilevazione di correlazioni tra le fasi del grooming (v. sez. 1.1) e l'arousal emotivo [45].

2.1 Natural Language Processing

Il Natural Language Processing (NLP) [46] è un ambito di ricerca interdisciplinare che si occupa dello studio del linguaggio umano elaborato da computer, integrando concetti di linguistica computazionale [47], Intelligenza

Artificiale (IA) [48] e Machine Learning [49], per la rilevazione di segnali psicologici e comportamentali di manipolazione, coercizione e abuso [50].

In particolare, nell'ambito del grooming, l'NLP consente di individuare segnali di abuso di fiducia, manipolazione emotiva, persuasione ingannevole e tentativi di controllo psicologico, anche quando tali segnali si manifestano in modo velato o indiretto. Le tecniche e modalità di analisi del linguaggio naturale sono state valutate e adattate per garantire precisione e affidabilità nel caso d'uso finale. I principali compiti che rientrano in questo ambito [51] includono nel nostro caso:

1. Classificazione di frasi
2. Classificazione delle singole parole in una frase
3. Generazione di contenuti testuali partendo da testi esistenti. Nel nostro lavoro, questa capacità è stata impiegata nella generazione di dati sintetici utili all'addestramento e alla validazione delle diverse componenti del modello di classificazione finale.

Il rilevamento di comportamenti manipolatori complessi richiede una preparazione accurata dei dati, sia nella raccolta dei dati di training che nel preprocessing [52] dei dati in input. Inoltre, la scelta delle tecniche di analisi, come l'uso di embedding semantici avanzati [43] o l'analisi contestuale delle relazioni tra le parole [53], è cruciale per ottimizzare la performance (i.e., risultati precisi) del modello.

Preprocessing

L'elaborazione del linguaggio naturale richiede l'applicazione di tecniche di preprocessing finalizzate a trasformare il testo libero in una rappresentazione strutturata e normalizzata (i.e., consiste nel trasformare il testo in una forma canonica e comparabile), adatta all'elaborazione automatica [54]. Tra le tecniche più comuni figurano la tokenizzazione, lo stemming, la lemmatizzazione [55] e gli embedding [56], ciascuna delle quali contribuisce

in modo specifico all’analisi linguistica.

La **tokenizzazione** è il primo passaggio nell’elaborazione del testo, che consiste nel suddividere un testo in unità più piccole, definite token, che possono rappresentare parole, frasi o simboli.

Lo **stemming** riduce le parole alla loro forma radicale, eliminando suffissi o prefissi e mantenendo la radice morfologica delle parole.

La **lemmatizzazione** consente nel preservare le relazioni tra le parole all’interno della frase, fondamentale specialmente quando il significato delle parole dipende dalla loro posizione e funzione grammaticale all’interno del contesto.

Gli **embedding** sono tecniche avanzate che rappresentano le parole come vettori numerici, catturando non solo il significato intrinseco delle parole, ma anche il contesto in cui esse appaiono. Questi vettori permettono di identificare le relazioni semantiche tra le parole, facilitando l’analisi e la comparazione tra termini. I modelli di embedding più noti, come *Word2Vec* [57], *GloVe* [58] e *FastText* [59], rappresentano le parole come vettori fissi, mentre i modelli più recenti, come *BERT* [60] e *GPT* [61], utilizzano embedding contestuali, che si adattano dinamicamente al contesto circostante, permettendo una rappresentazione più accurata e flessibile delle parole.

Modelli Distribuzionali e Contestuali

I modelli di rappresentazione del linguaggio naturale possono essere distinti in varie categorizzazioni, tra cui quella che distingue modelli distribuzionali [43] e contestuali [53]. Entrambi mirano a mappare le parole in uno spazio vettoriale continuo, in modo che la distanza o la direzione tra i vettori rifletta relazioni semantiche [44]. Tuttavia, differiscono nella modalità con cui trattano il contesto linguistico.

Modelli distribuzionali

I modelli distribuzionali si fondano sull’ipotesi secondo cui il significato di una parola è dato dai contesti in cui essa appare e apprendono rappresentazioni dense e a bassa dimensionalità, osservando le co-occorrenze statistiche delle parole in un ampio corpus testuale, deducendo dunque il significato sulla base della frequenza con cui una parola compare accanto ad altre. Tali modelli sono statici: ogni parola è associata ad un unico vettore, invariabile rispetto al contesto d’uso. Ciò implica, ad esempio, che parole polisemiche (i.e., stessa sintassi ma significato diverso a seconda del contesto) condividano la medesima rappresentazione numerica, rendendo il modello incapace di distinguere i diversi sensi lessicali in funzione dell’ambiente linguistico.

Tra i più noti modelli distribuzionali vi sono quelli basati su reti neurali predittive e su matrici di co-occorrenza. Nel primo gruppo rientra *Word2Vec*, che apprende gli embedding massimizzando la probabilità di predire una parola dato il contesto (architettura *CBOW* [62]) o viceversa (architettura *Skip-gram* [63]). Nel secondo gruppo, *GloVe* costruisce una matrice globale delle co-occorrenze tra parole, ottimizzando una funzione obiettivo che preserva proporzioni semantiche tra vettori. *FastText* estende *Word2Vec* rappresentando ciascuna parola come somma di vettori associati ai suoi n-grammi (i.e., sequenze di lunghezza n) di caratteri, migliorando la generalizzazione su vocaboli rari o non ancora visti. L’approccio distribuzionale risulta intrinsecamente limitato nell’analisi contestuale, indipendentemente dal tipo di modello utilizzato: la staticità delle rappresentazioni impedisce di cogliere le sfumature di significato che emergono da differenti contesti d’uso. Per ovviare a questa limitazione, sono stati introdotti i modelli contestuali.

Modelli contestuali

I modelli contestuali producono rappresentazioni dinamiche delle parole, condizionate dal contesto sintattico e semantico in cui si trovano, generando un embedding diverso per ogni occorrenza della stessa parola. Questa pro-

prietà consente di gestire la polisemia. Tali modelli si basano su architetture capaci di elaborare sequenze testuali e catturare dipendenze a lungo raggio. I primi approcci, come *ELMo* [64], impiegano reti *LSTM bidirezionali* [65] pre-addestrate su grandi corpora, con obiettivi linguistici generici, successivamente adattabili a task specifici. Con l'introduzione del meccanismo di *self-attention* [66] e dell'architettura *Transformer* [33], sono emersi modelli come Bidirectional Encoder Representations from Transformers (*BERT* [60]) e Generative Pre-trained Transformer (*GPT*) [61]. *BERT* adotta una codifica completamente bidirezionale, in cui ogni parola viene rappresentata considerando simultaneamente il contesto precedente e successivo. Questa caratteristica lo rende particolarmente adatto a compiti di comprensione del linguaggio, come la classificazione, il question answering [67] e il named entity recognition [68]. *GPT*, al contrario, utilizza una codifica unidirezionale (da sinistra a destra), privilegiando applicazioni generative, come la produzione autonoma di testo.

Emotion recognition

L'*emotion recognition* [69], pilastro fondamentale dell'affective computing [70], riguarda l'identificazione e l'interpretazione degli stati emotivi umani attraverso molteplici modalità, quali espressioni facciali o testuali, intonazioni vocali e segnali fisiologici. Questo campo interdisciplinare sfrutta tecniche avanzate di machine learning, in particolare il deep learning [71], per analizzare pattern complessi nei dati e migliorare l'accuratezza nel riconoscimento delle emozioni.

Le metodologie di riconoscimento si basano su modelli teorici strutturati, come il modello di Ekman [39], che identifica sei emozioni di base universalmente riconosciute e il modello a ruota di Plutchik [40], che organizza le emozioni in una struttura multidimensionale basata su intensità e relazioni tra emozioni [72]. Questi modelli forniscono un solido quadro concettuale per la progettazione e lo sviluppo di sistemi automatici di rilevamento emotivo.

Gli approcci più recenti privilegiano strategie multimodali [73], integrando segnali visivi, uditivi e contestuali per aumentare l'affidabilità e l'adattabilità in applicazioni reali, quali l'interazione uomo-computer, la robotica sociale e il monitoraggio della salute mentale. Studi pionieristici, come quelli di Picard [70], hanno introdotto il paradigma dell'affective computing, mentre Ekman [74] ha evidenziato l'universalità di alcune espressioni facciali, consolidando così le basi teoriche per sistemi automatizzati capaci di interpretare segnali emotivi in modo affidabile e generalizzabile.

Nel presente lavoro, l'*emotion recognition* sarà applicato all'analisi del testo scritto, con l'obiettivo di rilevare e classificare le espressioni emotive contenute nei dati, individuando pattern emotivi ricorrenti in specifiche sequenze testuali.

Semantic Similarity

La *Semantic Similarity* [44] esprime la vicinanza semantica tra due unità testuali, a prescindere dalla forma sintattica o lessicale. Nei modelli di linguaggio neurali, le frasi vengono rappresentate come vettori densi nello spazio \mathbb{R}^d , detti embedding (v. sez. 2.1), ottenuti mediante architetture pre-addestrate come *BERT* [32] e le sue variazioni. Il confronto tra vettori è basato su metriche di prossimità semantica (i.e., similarità o distanza) tra le quali, in questo studio, sono state analizzate:

- **Cosine similarity** [75]:

$$\text{sim}_{\cos}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \cdot \|\mathbf{v}_2\|}$$

dove $\mathbf{v}_1 \cdot \mathbf{v}_2$ è il prodotto scalare tra i vettori e $\|\mathbf{v}_1\|$ e $\|\mathbf{v}_2\|$ rappresentano la norma rispettivamente del vettore \mathbf{v}_1 e \mathbf{v}_2 . Nei modelli di embedding semantico che usiamo, l'informazione è codificata nella direzione dei

vettori. La cosine similarity valuta il coseno dell'angolo (i.e., l'angolo riflette l'orientamento) tra i vettori, risultando indipendente dalla loro norma. In questo modo, permette di rappresentare in maniera accurata la prossimità sintattica, oltre a quella semantica.

- **Distanza Euclidea (L2)** [76]:

$$d_{L2}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$$

con \mathbf{x}, \mathbf{y} vettori nello spazio \mathbb{R}^d , x_i e y_i rispettivamente la i -esima componente dei vettori \mathbf{x} e \mathbf{y} , d la dimensione dello spazio vettoriale e $\sum_{i=1}^d (x_i - y_i)^2$ la somma dei quadrati delle differenze tra le componenti corrispondenti dei due vettori. La distanza euclidea è stata considerata come metrica di similarità poiché incorpora sia l'informazione direzionale (orientamento) sia la componente scalare (norma) dei vettori nello spazio di embedding. Tuttavia, nei modelli di rappresentazione semantica, la norma dei vettori può risultare influenzata da fattori di training o da proprietà intrinseche del modello che non necessariamente riflettono la distanza semantica tra le frasi.

- **Jaccard Similarity** [77]:

$$\text{sim}_{\text{Jaccard}}(u, v) = \frac{\sum_i \min(u_i, v_i)}{\sum_i \max(u_i, v_i)}$$

Dove \mathbf{u} e \mathbf{v} sono vettori nello spazio \mathbb{R}^d , con d che rappresenta la dimensione dello spazio vettoriale. Le componenti u_i e v_i indicano rispettivamente la i -esima componente dei vettori \mathbf{u} e \mathbf{v} . Nel numeratore, $\sum_i \min(u_i, v_i)$ rappresenta la somma dei valori minimi tra le componenti corrispondenti dei due vettori, mentre il denominatore, $\sum_i \max(u_i, v_i)$, è la somma dei valori massimi delle stesse componenti. La similarità di Jaccard è stata utilizzata nel presente lavoro perché, elaborando il testo come insiemi di n-grammi, questa metrica permette di quantificare in modo diretto

e interpretabile la quota di elementi lessicali condivisi, fornendo un'indicazione immediata della prossimità semantica, indipendentemente da variazioni nella struttura sintattica o nella distribuzione vettoriale dei dati.

- **Dice Coefficient** [78]:

$$\text{sim}_{\text{Dice}}(u, v) = \frac{2 \sum_i \min(u_i, v_i)}{\sum_i u_i + \sum_i v_i}$$

Qui, \mathbf{u} e \mathbf{v} sono vettori in \mathbb{R}^d con componenti u_i e v_i . La formula misura la similarità considerando il doppio della somma dei minimi delle componenti, normalizzato dalla somma totale delle componenti di entrambi i vettori. Il coefficiente di Dice è stato utilizzato in questo lavoro poiché rappresenta una misura efficace per valutare la sovrapposizione relativa tra insiemi, risultando particolarmente adatto a dati basati su frequenze di elementi condivisi. Questa metrica, simile alla similarità di Jaccard ma con una maggiore sensibilità alla presenza comune, consente di quantificare con precisione la similarità tra rappresentazioni testuali, facilitando il confronto tra frasi in termini di condivisione lessicale.

- **Three-Way Jaccard (3W-Jaccard)** [79]:

$$\text{sim}_{\text{3WJ}}(u, v) = \frac{a}{a + b + c}$$

dove $a = \sum u_i v_i$ (componenti congiuntamente attive), $b = \sum u_i (1 - v_i)$ e $c = \sum (1 - u_i) v_i$ (componenti discordanti).

Il Three-Way Jaccard considera simultaneamente tre insiemi, permettendo di valutare la similarità tenendo conto non solo della frase di riferimento e della frase candidata, ma anche di un ulteriore insieme contestuale o informativo. Questo approccio consente di analizzare più frasi o elementi in modo congiunto, migliorando l'accuratezza nel rilevamento delle corri-

spondenze semantiche rispetto alla Jaccard classica, che confronta solo due insiemi.

- **Sokal-Sneath 1** [80]:

$$\text{sim}_{\text{SS1}}(u, v) = \frac{a}{a + 2b}$$

In questa formula, a rappresenta il numero di elementi comuni a entrambi i vettori o insiemi u e v , mentre b indica il numero di elementi presenti in uno solo dei due.

- **Sokal-Sneath 2** [81]:

$$\text{sim}_{\text{SS2}}(u, v) = \frac{2a}{2a + b}$$

con a che indica il numero di elementi condivisi tra i vettori o insiemi u e v , mentre b rappresenta il numero di elementi presenti in uno solo dei due. Questa misura enfatizza maggiormente gli elementi comuni rispetto a quelli unilaterali, quantificando la similarità in modo asimmetrico. I coefficienti di Sokal-Sneath 1 e 2 sono stati utilizzati in questo lavoro perché rappresentano metriche di similarità adatte per dati binari e categoriali, permettendo di valutare in modo distinto la presenza e l'assenza concordante degli attributi tra due insiemi (i.e., nel nostro caso frasi). In particolare, Sokal-Sneath 1 dà maggiore peso alle concordanze positive (presenza condivisa di elementi), mentre Sokal-Sneath 2 considera sia le concordanze positive sia le assenze condivise, offrendo così una valutazione più equilibrata della similarità. Queste caratteristiche li rendono utili nel contesto dell'analisi semantica basata su rappresentazioni discrete, come insiemi di n-grammi o caratteristiche lessicali, facilitando un confronto dettagliato tra frasi.

- **Roger-Tanimoto** [82]:

$$\text{sim}_{\text{RT}}(u, v) = \frac{a + c}{a + 2b + c}$$

Dove a rappresenta il numero di elementi attivi (o presenti) in entrambi i vettori u e v , b è il numero di elementi attivi in uno solo dei due vettori, mentre $c = \sum(1 - u_i)(1 - v_i)$ indica il numero di componenti congiuntamente inattive (i.e., elementi assenti in entrambi). Questa misura valuta la similarità tenendo conto sia delle presenze comuni che delle assenze comuni, fornendo una metrica bilanciata tra corrispondenze positive e negative. Questa rilevanza del concetto di assenza, è fondamentale nel caso della similarità testuale per velocizzare e migliorare la capacità di cogliere elementi di similarità o dissimilarità tra frasi, valorizzando anche informazioni di divergenza.

- **Faith Similarity** [83]:

$$\text{sim}_{\text{Faith}}(u, v) = a + 0.5b$$

dove a rappresenta il numero di elementi comuni a entrambi i vettori, mentre b indica il numero di elementi presenti in uno solo dei due.

La scelta di questa metrica è motivata dalla volontà di dare massimo peso agli elementi lessicali condivisi, che rappresentano una forte indicazione di similarità semantica, pur riconoscendo parzialmente la rilevanza anche degli elementi esclusivi. A differenza di metriche più drastiche (come Jaccard o Dice), la Faith Similarity attenua la penalizzazione delle differenze lessicali, offrendo una misura più tollerante e sfumata della similarità tra frasi, particolarmente utile nel confronto tra le rappresentazioni discrete delle frasi, caratterizzate da alta varianza di strutture e espressioni linguistiche, da considerare in modo preciso e sottile.

Queste metriche sono state considerate al fine di analizzare come diverse definizioni di prossimità influenzino la valutazione semantica tra frasi rappresentate in vettori, sia densi che binari. Ciascuna metrica presenta vantaggi

teorici e limiti a seconda della natura della rappresentazione e dell’obiettivo semantico perseguito. La ricerca della frase più simile a partire da un vettore viene resa efficiente attraverso strutture per *Approximate Nearest Neighbor Search* (ANNS) [84]. La libreria *FAISS* [85], in particolare, consente l’indicizzazione e la ricerca veloce in grandi spazi vettoriali tramite strategie come l’*inverted index* (*IVF*) [86], la *quantizzazione a prodotto* (*PQ*) [87] e metodi esatti come *IndexFlatL2* [88].

Valutazione delle metriche di prossimità semantica

Per valutare l’efficacia di diverse metriche di prossimità semantica per la classificazione delle emozioni in un data set bilanciato di frasi (v. sez. 2.5), abbiamo condotto uno studio preliminare, su embedding generati da sette modelli differenti di Sentence Transformers [89]. L’obiettivo è stato isolare l’effetto della metrica, indipendentemente dalle caratteristiche specifiche dei modelli: infatti, se più modelli ottengono i migliori risultati con la stessa metrica, si può dedurre che la bontà delle performance sia attribuibile alla metrica stessa, e non al singolo modello.

La classificazione è avvenuta calcolando la similarità tra gli embedding delle frasi e determinando l’etichetta più frequente tra i k vicini più simili (metodo *k-nearest neighbors*) [90].

Sia f_{test} una frase del set, rappresentata dal suo embedding vettoriale \vec{f}_{test} . L’insieme dei k vicini più simili nel data set di training è definito come:

$$\mathcal{N}_k(f_{\text{test}}) = \arg \max_{f \in D_{\text{train}}} \text{sim}(\vec{f}_{\text{test}}, \vec{f}) \quad (2.1)$$

dove sim rappresenta una delle misure di similarità. L’etichetta finale assegnata alla frase di test è determinata dalla modalità delle etichette dei vicini:

$$\hat{y}(f_{\text{test}}) = \text{mode}(\{y(f) \mid f \in \mathcal{N}_k(f_{\text{test}})\}) \quad (2.2)$$

I risultati presentati in Tabella 2.2, giustificano la scelta della *Cosine Similarity* (v. sez. 2.1), in quanto si è dimostrata, in maniera consistente, la metrica con le prestazioni migliori.

Notiamo che anche la metrica di *Roger Tanimoto* (v. sez. 2.1) ha mostrato buone performance per alcuni modelli, equiparabili alla *Cosine Similarity*. Per questo motivo, verrà presa in considerazione come potenziale alternativa tra le metriche analizzate. I valori di accuratezza (ottenuta confrontando se la label predetta e label effettiva nel data set coincidono) riportati in tabella sono esclusivamente quelli relativi a $k = 1, 3, 5$, poiché un aumento a $k = 9$ comporta un decremento delle prestazioni in termini di accuratezza, probabilmente dovuto alla distanza eccessiva dei vicini considerati. Nella tabella 2.1, si mostrano a titolo dimostrativo alcuni esempi di tale andamento.

Modello	Metrica	k=1	k=3	k=5	k=9
allMini	jaccard	0.2170	0.2220	0.2378	0.2138
all_MPNet	cosine	0.8115	0.8125	0.8128	0.8088
ML_para	sokal_sneath_1	0.0993	0.0990	0.0863	0.0840

Tabella 2.1: Esempi di metriche con prestazione peggiore per $k = 9$ rispetto a $k = 1, 3, 5$

Tale analisi ha evidenziato un degrado uniforme delle prestazioni per valori di $k > 5$, indipendentemente dal modello considerato; pertanto, tali valori sono stati omessi nei successivi confronti.

Tabella 2.2: Accuratezze delle metriche di similarità per diversi modelli (valori più alti in grassetto)

Metrica	k=1	k=3	k=5
multilingual-paraphrase-mpnet			
cosine	0.7937	0.7930	0.7993

Continua nella pagina successiva

Tabella 2.2 – continuazione

jaccard	0.4660	0.4735	0.4572
dice	0.4318	0.4345	0.4550
3w-jaccard	0.1308	0.1330	0.1353
sokal_sneath_1	0.0993	0.0990	0.0862
sokal_sneath_2	0.2585	0.2925	0.3513
roger_tanimoto	0.7650	0.7750	0.7810
faith	0.1250	0.1250	0.1250
<hr/>			
roberta			
cosine	0.8127	0.8187	0.8203
jaccard	0.4233	0.4010	0.3815
dice	0.3290	0.3633	0.4153
3w-jaccard	0.4285	0.4798	0.5232
sokal_sneath_1	0.3132	0.3362	0.3840
sokal_sneath_2	0.3967	0.5128	0.5952
roger_tanimoto	0.7857	0.7947	0.8007
faith	0.1250	0.1250	0.1250
<hr/>			
all-MiniLM-L6-v2			
cosine	0.7378	0.7462	0.7542
jaccard	0.2170	0.2220	0.2377
dice	0.1242	0.1263	0.1390
3w-jaccard	0.1775	0.1885	0.1940
sokal_sneath_1	0.1693	0.1715	0.1850
sokal_sneath_2	0.1507	0.1610	0.1815
roger_tanimoto	0.6122	0.6302	0.6445
faith	0.1247	0.1247	0.1250
<hr/>			
all-MPNet-base-v2			
cosine	0.8115	0.8125	0.8127
<hr/>			

Continua nella pagina successiva

Tabella 2.2 – continuazione

jaccard	0.1680	0.1678	0.1668
dice	0.1250	0.1250	0.1250
3w-jaccard	0.1123	0.1108	0.1072
sokal_sneath_1	0.0995	0.0983	0.1045
sokal_sneath_2	0.1730	0.1780	0.1968
roger_tanimoto	0.7750	0.7775	0.7805
faith	0.1247	0.1247	0.1245
paraphrase-distilroberta-base-v1			
cosine	0.8020	0.8135	0.8220
jaccard	0.6315	0.6285	0.6242
dice	0.6315	0.6285	0.6242
3w-jaccard	0.3733	0.4113	0.4615
sokal_sneath_1	0.2233	0.2320	0.2547
sokal_sneath_2	0.7575	0.7698	0.7907
roger_tanimoto	0.7880	0.7957	0.8030
faith	0.1247	0.1247	0.1250
multi-qa-mpnet-base-dot-v1			
cosine	0.8170	0.8265	0.8325
jaccard	0.8210	0.8267	0.8230
dice	0.8210	0.8267	0.8230
3w-jaccard	0.1710	0.1797	0.1915
sokal_sneath_1	0.0752	0.0703	0.0607
sokal_sneath_2	0.8085	0.8255	0.8257
roger_tanimoto	0.8225	0.8277	0.8383
faith	0.1247	0.1247	0.1245
sentence-t5-large			
cosine	0.8498	0.8515	0.8555

Continua nella pagina successiva

Tabella 2.2 – continuazione

jaccard	0.1870	0.1450	0.1420
dice	0.1250	0.1250	0.1250
3w-jaccard	0.1492	0.1467	0.1465
sokal_sneath_1	0.1395	0.1427	0.1455
sokal_sneath_2	0.1383	0.1355	0.1353
roger_tanimoto	0.1253	0.1205	0.2003
faith	0.1247	0.1250	0.1242

2.2 Classificazione del linguaggio in classi emotive

Tra gli approcci classici alla classificazione, si possono individuare le due categorie dei sistemi *rule-based* [91] e basati su *deep learning* [31].

I metodi rule-based si basano su regole esplicite, derivate da pattern predefiniti, che fanno uso di liste lessicali composte da termini e locuzioni con una specifica accezione nell’ambito di studio. Nel contesto specifico della nostra ricerca, tali tecniche potrebbero essere utilizzate per il riconoscimento di una determinata emozione, attraverso l’identificazione di vocaboli semanticamente correlati a quella particolare dimensione emotiva. Formalmente, ogni regola r_i definisce un pattern da ricercare nel testo T , e l’output complessivo è la combinazione logica di tali regole:

$$r_i(T) = \begin{cases} 1, & \text{se } T \text{ soddisfa il pattern } i \\ 0, & \text{altrimenti} \end{cases} \Rightarrow f_{\text{rule}}(T) = \bigvee_{i=1}^m r_i(T).$$

Questi sistemi eccellono per essere interpretabili e giustificabili in ogni dettaglio, ma risultano complessi da implementare e da adattare a variazioni non previste, limitando la capacità di rilevare emozioni espresse nel testo mediante forme non convenzionali.

Modelli di Machine Learning

I modelli di Machine Learning (ML) [49] apprendono direttamente dai dati che vorremmo in output, permettendo l’adattamento a nuove espressioni e pattern linguistici, permettendo l’addestramento dai dati che vogliamo in output (ossia dando in training le label associate ai dati), senza necessità di definire regole. Questi modelli risultano particolarmente efficaci nell’analisi di grandi corpora testuali, nei quali è impraticabile definire a priori tutte le possibili varianti del linguaggio. Tra i principali algoritmi utilizzati per la classificazione del linguaggio si annoverano:

- **Naïve Bayes** [92]: classificatore probabilistico basato sull’assunzione di indipendenza condizionale tra le feature testuali. La classificazione si fonda sul calcolo della probabilità a posteriori tramite il teorema di Bayes [93]:

$$P(C | X) = \frac{P(X | C)P(C)}{P(X)}$$

dove C è la classe e X il vettore delle caratteristiche. Indipendentemente dalla semplicità e dall’efficienza computazionale che caratterizzano questo algoritmo, il classificatore *Naive bayes* risulta meno efficace nel trattamento del linguaggio naturale perchè assume che ogni parola contribuisca in modo indipendente al significato del testo [94]. Questa assunzione, spesso non realistica nell’ambito NLP [46], limita la capacità del modello nel cogliere relazioni all’interno delle frasi, che sono invece fondamentali per comprendere correttamente il contesto e rilevare correttamente le emozioni.

- **Support Vector Machines (SVM)** [95]: modelli discriminativi che individuano un iperpiano $H : \mathbf{w} \cdot \mathbf{x} + b = 0$ massimizzando il margine tra le classi. Sono robusti nella classificazione binaria e particolarmente adatti a spazi ad alta dimensionalità. Le *SVM* sono efficaci nel distinguere pattern complessi di linguaggio, in quanto riescono a separare in modo preciso diverse categorie di testo, anche quando i dati non sono linearmente

separabili. Tuttavia, le *SVM* non tengono conto dell’ordine delle parole né delle relazioni contestuali tra di esse [96].

- **Random Forest** [97]: ensemble di alberi decisionali che aggrega predizioni di molteplici modelli deboli per ridurre varianza [98] e overfitting [99] e mantenere la explainability (i.e., quanto facilmente possiamo capire e giustificare il funzionamento di un modello o le sue decisioni), con la possibilità di verificare le condizioni di ogni scelta negli alberi di decisione. La classificazione si basa su una votazione a maggioranza delle decisioni individuali:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\}$$

dove h_t rappresenta l’albero decisionale t -esimo e T il numero totale degli alberi. Questo modello è efficace nell’identificare schemi complessi e non lineari.

Nonostante la maggiore adattabilità rispetto ai sistemi rule-based [91], i modelli ML tradizionali presentano limitazioni legate alla necessità di data set di dimensioni rilevanti e potenza di calcolo elevata, per una corretta generalizzazione in fase di addestramento, oltre alla dipendenza critica da qualità, bilanciamento e rappresentatività dei dati [100]. D’altra parte, il risultato finale è ottenuto combinando gli outputs di più classificatori, con l’obiettivo di massimizzare la capacità di generalizzazione e di evitare le distorsioni, le allucinazioni (i.e., fenomeno in cui il modello genera contenuti che sembrano plausibili ma sono falsi o inventati) o i limiti di ciascun modello considerato singolarmente [101].

Deep Learning

Gli approcci basati su Deep Learning [71] hanno buone performance per l’analisi e la classificazione del linguaggio, grazie alla capacità di apprendere rappresentazioni gerarchiche e distribuite dei dati testuali, captando segnali

sottili spesso impercettibili a modelli tradizionali. Tra i modelli più noti si distinguono:

- **RNN (Recurrent Neural Networks)** [102]: reti neurali ricorrenti progettate per il trattamento di sequenze temporali, con dinamiche interne che permettono di incorporare informazioni del contesto precedente. La loro struttura si basa su equazioni ricorsive del tipo:

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

dove h_t è lo stato nascosto al tempo t , x_t l'input corrente e σ una funzione di attivazione. Nonostante la capacità di modellare dipendenze sequenziali, le RNN soffrono del problema del *vanishing gradient* [103], limitando l'apprendimento su sequenze lunghe.

- **LSTM (Long Short-Term Memory)** [65]: estensione delle *RNN* che introduce meccanismi di gating [104] per mitigare il vanishing gradient, consentendo la memorizzazione e il recupero di informazioni su periodi temporali prolungati. Il core dell'*LSTM* è descritto da:

$$\begin{cases} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{cases}$$

Nonostante miglioramenti rispetto alle *RNN* tradizionali, le *LSTM* mantengono un'elaborazione sequenziale, senza parallelizzazione e con efficienza computazionale limitata su dati di grandi dimensioni.

- **Transformer** [33]: architettura basata sul meccanismo di *self-attention* [66], che consente la modellazione delle dipendenze globali nel testo trattando simultaneamente tutte le posizioni della sequenza. La componente

centrale, l'attenzione scalata, si definisce come:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

dove Q, K, V sono rispettivamente query, key e value derivati dagli input, e d_k è la dimensione delle key. Modelli come *BERT* (Bidirectional Encoder Representations from Transformers) [60] e *GPT* (Generative Pretrained Transformer) [61] si basano su questa architettura, superando *RNN* e *LSTM* sia in termini di capacità di catturare dipendenze a lungo raggio sia di efficienza computazionale.

L'addestramento di tali modelli richiede elevati costi computazionali e ingenti quantità di dati. Tuttavia, tecniche avanzate di fine-tuning [105] come il Transfer Learning [106] consentono di adattare i modelli, una volta pre-addestrati su ampi corpora generici, a compiti di classificazione addizionali su set più specifici di classi (e.g., le emozioni), utilizzando relativi data set specifici per l'addestramento di tali classi, che possono essere significativamente più piccoli rispetto a quelli necessari all'addestramento iniziale.

Formalmente, dato un modello pre-addestrato con parametri θ_0 , questa tecnica ottimizza una funzione di *loss* \mathcal{L} [107] su un data set specifico:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\theta; \mathcal{D}_{task})$$

partendo da θ_0 , preservando così le conoscenze generali acquisite (e.g., congelando i pesi dei primi livelli di addestramento) e specializzandole sul task (e.g., riaddestrando gli ultimi livelli, deputati al calcolo della probabilità sul set di classi o aggiungendo livelli aggiuntivi ad hoc).

I vantaggi includono:

- Riduzione significativa del costo computazionale e della quantità di dati necessari rispetto al riaddestramento da zero;
- Miglioramento dell'accuratezza sul task specifico grazie all'adattamento contestuale;

- Maggiore flessibilità nell’aggiornamento e riadattamento a nuovi dati o task correlati, evitando il problema di catastrophic forgetting [108], cioè la tendenza di un modello a dimenticare le conoscenze pregresse, sovrascrivendole totalmente con le nuove informazioni apprese.

Settings

Per ottimizzare il fine-tuning [105] di un modello pre-addestrato, sia in termini di costi computazionali che di bontà dei risultati (i.e., valutati rispetto alle metriche descritte nella sottosezione 2.2), le principali tecniche, sfuttate anche in questa ricerca, sono:

- **Congelamento dei Layer:** in questo approccio [109], i primi strati del modello pre-addestrato vengono "congelati" (ovvero non aggiornati durante il fine-tuning) e solo gli strati superiori vengono allenati. Questo consente di preservare le conoscenze generali apprese dal modello durante il pre-addestramento, concentrandosi sull’adattamento ai nuovi dati.
- **Apprendimento a Tasso Variabile:** il tasso di apprendimento [110] viene inizialmente aumentato gradualmente (fase di *warm-up* [111]) per poi essere ridotto. Questa strategia aiuta a migliorare la stabilità dell’addestramento e a evitare che il modello si "adatti troppo rapidamente" durante le prime fasi dell’addestramento, prevenendo l’instabilità nei risultati.
- **Early Stopping:** questa tecnica [112] monitora le prestazioni del modello sulla set di validazione e interrompe l’addestramento quando non si osservano miglioramenti significativi, prevenendo l’overfitting e risparmiando risorse computazionali.

Nel presente studio, tale approccio è stato sfruttato nell’addestramento dei classificatori preposti al rilevamento delle emozioni primarie e secondarie all’interno dei testi analizzati. Altre metodologie avanzate utilizzate al fine di adattare i modelli al task specifico sono:

- **In-Context Learning** [113]: apprendimento contestuale tramite esempi forniti direttamente nel prompt, senza necessità di aggiornare i pesi del modello. Questo approccio si è rivelato fondamentale nella fase di generazione di frasi aderenti ad un’emozione specifica per popolare il data set (v. sez. 2.5), garantendo la flessibilità e possibilità di variare il tipo di output senza apportare modifiche costose dal punto di vista computazionale, al modello.
- **RAG** [41]: integra il recupero durante l’inferenza (i.e., cioè il processo mediante il quale il modello genera risposte partendo dall’input fornитогли) di informazioni esterne da risorse specifiche, migliorando il grado di personalizzazione e di conoscenza del modello preaddestrato. Il RAG è stato impiegato nella generazione del data set descritto nella sezione 3.5.
- **Prompting** [114]: costituisce la progettazione strategica degli input testuali per guidare il comportamento del modello verso risposte o classificazioni più precise e specifiche al task; è stato utilizzato in modo complementare all’ *In-Context Learning*, per massimizzare la precisione e la bontà dei risultati ottenuti durante il processo di *inferenza*.

Metriche di valutazione

Le prestazioni di ciascun modello dopo l’addestramento (i.e., fase di training) sul task specifico, vengono analizzate utilizzando metriche appropriate, che possono variare a seconda del tipo di obiettivo e delle caratteristiche del data set. Le seguenti metriche sono state utilizzate per ottenere una valutazione accurata delle performance di classificazione:

- **F1 Score** [115]: la media armonica tra *precisione* [116] e *richiamo* [117], utile in scenari con data set sbilanciati, poiché considera sia la capacità del modello di identificare correttamente le classi positive che la sua capacità di non produrre falsi positivi.
- **AUC-ROC (Area Under the Curve)** [118]: la curva *ROC* misura la

capacità del modello di distinguere tra classi positive e negative, bilanciando i falsi positivi e i veri positivi. L'area sotto la curva (*AUC*) [119] fornisce una valutazione complessiva delle prestazioni del modello.

- **Confusion Matrix** [120]: una matrice di confusione mostra i veri positivi, veri negativi, falsi positivi e falsi negativi, fornendo una visione chiara degli errori di classificazione e della distribuzione delle previsioni rispetto alle etichette reali.
- **Accuracy** [121]: Sebbene l'accuratezza possa non essere la metrica più adatta in scenari sbilanciati, rimane comunque una misura fondamentale per valutare la percentuale complessiva di previsioni corrette.

2.3 Large Language Model

I Large Language Model rappresentano una delle innovazioni più significative nell'elaborazione del linguaggio naturale. Questi modelli sono addestrati su enormi quantità di dati testuali provenienti da domini e contesti diversi, permettendo di catturare schemi semantici e sintattici complessi. Grazie alla loro architettura, sono in grado di modellare efficacemente le relazioni contestuali tra le parole in frasi e conversazioni, migliorando la classificazione e la generazione del testo.

Architettura Transformer

L'architettura Transformer è la base degli LLM moderni come *BERT* [60], *RoBERTa* [122], *GPT* [61], *T5* [123] e *Sentence Transformers* [89]. L'architettura *Transformer* [33] è composta da due componenti principali [124]:

- **Encoder:** trasforma la sequenza di input in una rappresentazione continua, catturando il significato contestuale.

- **Decoder:** genera una sequenza di output a partire dalla rappresentazione dell'encoder, utile per compiti di generazione testuale.

A seconda del modello, si utilizzano solo encoder (*BERT*, *RoBERTa*) [125], solo decoder (*GPT*) [126] o entrambi (*T5*, *BART*) [124].

Encoder

L'encoder riceve in input una sequenza di token $\mathbf{x} = (x_1, x_2, \dots, x_n)$ e produce una sequenza di vettori di rappresentazione $\mathbf{h} = (h_1, h_2, \dots, h_n)$, dove ogni $h_i \in \mathbb{R}^d$ cattura il significato contestuale del token x_i . Nei modelli come *BERT*, l'encoder è bidirezionale, cioè considera contemporaneamente il contesto a sinistra e a destra di ogni token, consentendo una comprensione più completa. Ogni strato di encoder contiene due componenti principali:

1. **Meccanismo di auto-attenzione (Self-Attention) [66]:** calcola una matrice di attenzione che pesa ogni parola in relazione a tutte le altre nella sequenza, formalmente definita come:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

dove Q, K, V sono rispettivamente le matrici di query, key e value derivate dagli input, e d_k è la dimensione delle key, usata per normalizzare il prodotto scalare [31]. Questo meccanismo consente al modello di catturare dipendenze a lungo raggio ed elaborare simultaneamente l'intera sequenza.

2. **Rete Neurale Feed-Forward (FFN) [127]:** dopo l'auto-attenzione, ogni token viene ulteriormente trasformato da una rete feed-forward applicata posizione per posizione:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

dove W_1, W_2 e b_1, b_2 sono parametri appresi e la funzione $\max(0, \cdot)$ rappresenta la *ReLU* [128].

Decoder

Il decoder è presente nei modelli autoregressivi e encoder-decoder, come *GPT* [61] e *T5* [123], che generano sequenze testuali. Riceve come input la rappresentazione dell'encoder e genera token uno alla volta, condizionando la predizione del token successivo sui token già generati. Nel decoder si distinguono due meccanismi di attenzione:

- **Auto-attenzione mascherata [129]:** simile all'encoder, ma vincolata a considerare solo i token precedenti nella sequenza di output, garantendo la generazione autoregressiva. La maschera impedisce l'accesso ai token futuri, assicurando che:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + M \right) V$$

dove M è una matrice di maschera che assegna $-\infty$ ai token futuri.

- **Attenzione incrociata (Cross-Attention) [130]:** Permette al decoder di focalizzarsi sulle rappresentazioni dell'encoder, integrando il contesto dell'input durante la generazione:

$$\text{Attention}(Q_d, K_e, V_e) = \text{softmax} \left(\frac{Q_d K_e^\top}{\sqrt{d_k}} \right) V_e$$

dove Q_d sono le query del decoder, mentre K_e, V_e sono key e value dell'encoder.

Questa combinazione consente ai modelli encoder-decoder di generare testi coerenti e contestualmente appropriati, risultando efficaci in compiti come la traduzione automatica, il completamento testuale e la risposta a domande.

Sentence Transformers

I Sentence Transformers [89] sono modelli avanzati nell'ambito dell'elaborazione del linguaggio naturale, progettati per generare rappresentazioni vettoriali a lunghezza fissa di intere frasi, anziché di singole parole. Questi

modelli, derivati e ottimizzati a partire dalle architetture *Transformer* [33], mirano a catturare il significato semantico complessivo di una frase, superando le limitazioni dei modelli che analizzano il testo a livello di token isolati. Formalmente, dato un input testuale costituito da una sequenza di token

$$\mathbf{x} = (x_1, x_2, \dots, x_n),$$

un Sentence Transformer produce un vettore di embedding 2.1

$$\mathbf{e} \in \mathbb{R}^d,$$

di dimensione fissa d , che sintetizza il contenuto semantico globale della frase. Per ottenere questa rappresentazione compatta, i modelli impiegano livelli di *pooling* [131] che aggregano le rappresentazioni contestuali token-level generate dall'encoder *Transformer*. In particolare, l'encoder produce una sequenza di vettori

$$\mathbf{H} = (h_1, h_2, \dots, h_n), \quad h_i \in \mathbb{R}^d,$$

dove ogni vettore h_i cattura il significato contestuale del token x_i . Il vettore di embedding finale \mathbf{e} viene calcolato tramite funzioni di pooling come:

$$\mathbf{e} = \text{pool}(\mathbf{H}) = \frac{1}{n} \sum_{i=1}^n h_i \quad (\text{mean pooling})$$

oppure mediante max pooling o l'estrazione del vettore corrispondente al token [CLS] [132].

La rappresentazione vettoriale a lunghezza fissa consente ai *Sentence Transformers* [89] di essere impiegati in numerosi compiti avanzati di NLP (v. sez. 2.1), quali:

- la misurazione della similarità semantica tra frasi [44],
- il recupero delle informazioni (information retrieval) [133],
- la classificazione del testo,
- il question answering [67],

- la paraphrase detection [134].

La capacità di confrontare efficacemente il significato di frasi anche in contesti complessi o multilingue rappresenta una delle principali caratteristiche distintive di questi modelli.

Architettura

I Sentence Transformers [89] modificano l’architettura standard basata su Transformer [33] per ottimizzare la generazione di embedding [56] a livello di frase. In particolare, adottano strutture di rete quali le Siamese Networks [135] e le Triplet Networks [136], che migliorano l’efficienza nell’apprendimento di rappresentazioni semantiche coerenti. Le caratteristiche architettoniche principali sono:

- **Architettura Siamese e Triplet Networks:** il modello viene addestrato minimizzando la distanza tra vettori embedding di frasi semanticamente simili e massimizzando quella tra frasi dissimili. Data una coppia di frasi (s_a, s_p) semanticamente simili e una frase negativa s_n , il modello ottimizza la seguente funzione di perdita triplet:

$$\mathcal{L} = \max(0, d(\mathbf{e}_a, \mathbf{e}_p) - d(\mathbf{e}_a, \mathbf{e}_n) + m),$$

dove $\mathbf{e}_a, \mathbf{e}_p, \mathbf{e}_n \in \mathbb{R}^d$ sono gli embedding delle rispettive frasi, $d(\cdot, \cdot)$ è una misura di distanza (tipicamente la distanza euclidea o la distanza coseno), e $m > 0$ è un margine predefinito. Questo approccio costruisce uno spazio di rappresentazione semantica in cui frasi con significati affini risultano vicine.

- **Dati di addestramento basati su inferenza logica** [137]: l’addestramento si basa su coppie di frasi etichettate come semanticamente simili o dissimili. Ciò consente al modello di apprendere le relazioni semantiche in un contesto logico, migliorando la capacità di discriminare significati affini da quelli divergenti.

Modelli Principali di Sentence Transformers

I principali modelli di Sentence Transformers [89] includono diverse varianti, ciascuna ottimizzata in funzione delle esigenze specifiche riguardo performance, efficienza computazionale e capacità multilingue. Tra questi, Sentence-BERT (*SBERT*) [138] è progettato per massimizzare la qualità nella misurazione della similarità semantica tra frasi, mentre modelli più leggeri come *DistilBERT* [139] e *MiniLM* [140] offrono un buon compromesso tra accuratezza e complessità computazionale. *XLM-R* [141] si distingue per applicazioni multilingue, mentre *RoBERTa* [122], *ALBERT* [142] e *MPNet* [143] si focalizzano su miglioramenti prestazionali per attività quali classificazione, question answering e rilevamento di parafrasi (v. sez. 2.1)

La Tabella 2.3 riassume le specifiche principali di questi modelli, inclusi dimensioni, numero di parametri e data set di addestramento.

Tabella 2.3: Confronto tra diversi modelli di linguaggio

Modello	Dimensione (embedding)	Numero di Parametri	Data set di Addestramento
SBERT (Sentence-BERT)	768	Circa 110 milioni	1.500.000 frasi per STS, Quora, e NLI
DistilBERT	768	Circa 66 milioni	Wikipedia (2.5B parole), BookCorpus (800M parole), CC-News (1B parole)
MiniLM	256 o 512	22 milioni	SNLI (570.000 frasi), MultiNLI (433.000 frasi)
XLM-R (Cross-Lingual RoBERTa)	768	550 milioni	2.5 TB di dati (lingue multiple)
RoBERTa	768	355 milioni	160 GB di testo (BookCorpus, Wikipedia, OpenWebText, CC-News)
ALBERT (A Lite BERT)	768	Circa 12 milioni	BookCorpus (800M parole), Wikipedia (2.5B parole)
MPNet	768	300 milioni	MS MARCO (8.8M documenti), SWAG (500.000 frasi), BookCorpus, WikiText-103
DistilUSE (Universal Sentence Encoder)	512	100 milioni	Wikipedia (2.5B parole), Stanford Sentiment Treebank (11.000 frasi), SICK (10.000 frasi)

Ad esempio, il modello all-roberta-large-v1 [122] conta circa 355 milioni

di parametri e presenta un’architettura composta da 24 strati transformer, ciascuno dei quali integra un modulo di multi-head self-attention [66] e un livello feed-forward [127] completamente connesso. Ogni strato è dotato di 16 teste di attenzione, per un totale di 384 canali di attenzione parallela, che consentono al modello di cogliere con efficacia le relazioni semantiche latenti tra i token all’interno della sequenza di input. La dimensione dell’embedding [56] è pari a 1024, offrendo uno spazio vettoriale sufficientemente ampio per rappresentare informazioni sintattiche e semantiche complesse. Il vocabolario utilizzato è stato costruito mediante il metodo di tokenizzazione Byte-Pair Encoding (BPE) [144] e comprende 50.265 token distinti. Un token speciale, [CLS] [132], è inserito all’inizio di ogni sequenza e viene impiegato per aggregare le informazioni globali del testo, fornendo l’input alle teste di classificazione per le previsioni finali.

Per valutare l’efficacia di queste architetture, è stato condotto un esperimento preliminare volto a confrontare la similarità semantica calcolata tra coppie di frasi con differenze lessicali ma significati simili e con parole simili con significati diversi, utilizzando sia modelli basati su Transformer tradizionali (*BERT*) [60] sia su Sentence Transformers (*SBERT*) [138]. Per ogni coppia, sono stati calcolati gli embedding di frase tramite i diversi modelli, quindi è stata misurata la cosine similarity [75] tra i vettori risultanti. I risultati evidenziano come i modelli della famiglia *BERT* (*BERT-base*, *BERT-large* [32] e *RoBERTa* [122]) tendano a fornire valori elevati di similarità anche per coppie con significati semanticamente divergenti, suggerendo una dipendenza da segnali superficiali quali la presenza di termini comuni o similarità lessicale, piuttosto che da una rappresentazione semantica profonda. Al contrario, i modelli *Sentence-BERT* (i.e., *SBERT*) mostrano una maggiore capacità di disambiguazione, restituendo valori più coerenti con la reale affinità concettuale tra frasi.

2.4 Modello delle Emozioni di Plutchik

Il modello delle emozioni di Robert Plutchik [40], sviluppato nel 1980, è una delle teorie psicologiche più influenti nello studio delle emozioni umane. Le emozioni sono reazioni adattive ed evolutive che svolgono un ruolo fondamentale nella sopravvivenza dell'individuo, facilitando risposte appropriate a stimoli ambientali. In questo contesto, il modello di Plutchik è noto come "ruota" o "fiore" delle emozioni, una rappresentazione diagrammatica che organizza le *emozioni primarie* (che derivano dal più semplice modello di Paul Ekman [39]), le possibili intensità e le varie combinazioni, che definiamo *emozioni secondarie* e gli stati affettivi intermedi tra emozioni, in una visualizzazione intuitiva delle loro interrelazioni. La ruota, mostrata in figura 2.1, illustra come le emozioni possano variare in intensità, passando da forme più deboli (in basso nel modello 3D) a espressioni più intense (in alto nel modello 3D), rappresentando le sfumature emotive che caratterizzano l'esperienza umana. Nella ruota delle emozioni, ogni emozione primaria è rappresentata come un petalo della ruota e ogni petalo descrive tre intensità emotive. Inoltre, la posizione dei petali nel modello rappresenta emozioni simili (petali vicini) o, viceversa, opposte (petali in posizioni opposte).

Tale modello verrà utilizzato nel nostro lavoro come base strutturata e gerarchica per il rilevamento delle emozioni.

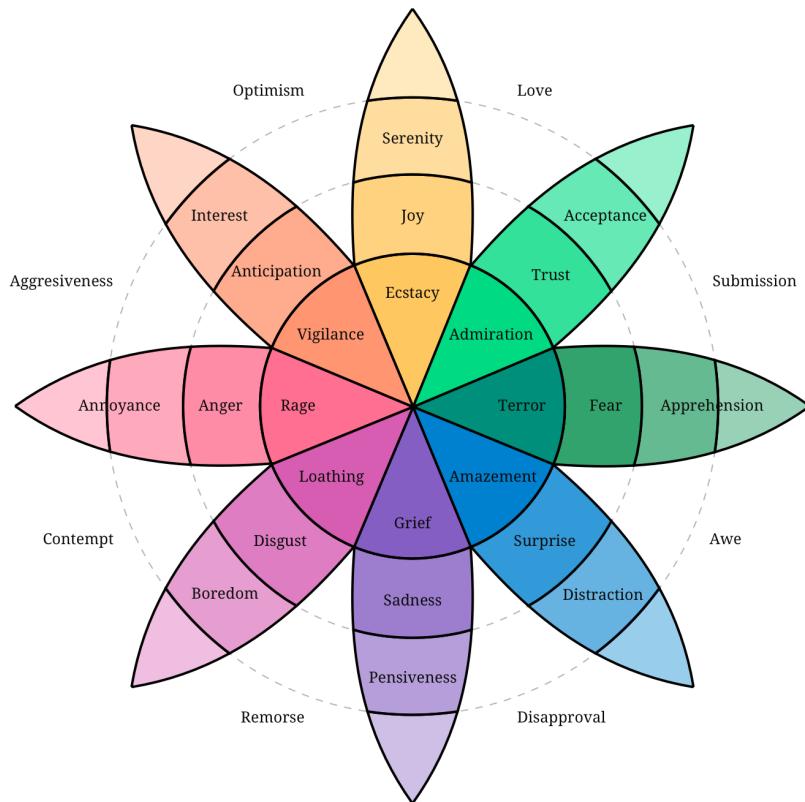


Figura 2.1: Modello delle emozioni di Plutchik

Il Modello di Plutchik e il Modello di Ekman

Il modello di Plutchik parte dal modello di Ekman, strutturato da Paul Ekman a partire da uno studio geografico e culturale sistematico da lui condotto sulle espressioni facciali, per identificare emozioni universali (i.e., di base) che in tutto il globo vengono riconosciute allo stesso modo. Ekman identificò inizialmente sei emozioni universali: *Joy* (gioia), *Sadness* (tristezza), *Fear* (paura), *Anger* (rabbia), *Surprise* (sorpresa) e *Disgust* (disgusto), estendendo poi il modello di base, con l'aggiunta di *Trust* (fiducia) e *Anticipation* (aspettativa). A partire dagli studi di Ekman, il modello di Plutchik offre una visione più articolata delle emozioni, proponendo le variazioni delle otto emozioni primarie nella dimensione dell'intensità.

Di seguito sono riportate le varianti di intensità per ciascuna delle emozioni primarie di Plutchik:

- **Joy (Gioia):**

- *Serenity*: una forma lieve di gioia, caratterizzata da una sensazione di piacere che non è travolgente.
- *Joy*: una felicità genuina che porta a un buon stato d'animo.
- *Ecstasy*: una gioia intensa e travolgente, accompagnata da una sensazione di benessere estremo.

- **Trust (Fiducia):**

- *Acceptance*: una forma moderata di fiducia, che implica una sensazione di stabilità e affidabilità.
- *Trust*: una convinzione solida in qualcosa o qualcuno, con una base razionale di sicurezza.
- *Admiration*: una forma intensa di fiducia, caratterizzata da una visione idealizzata della persona o della situazione.

- **Fear (Paura):**

- *Apprehension*: una forma lieve di paura, spesso associata a preoccupazioni future o incertezze.
- *Fear*: una reazione emotiva più forte a una minaccia percepita, che induce alla difesa o alla fuga.
- *Terror*: una paura intensa e paralizzante, che può portare a una reazione di panico.

- **Surprise (Sorpresa):**

- *Distraction*: una reazione di breve durata a un evento inaspettato, che può essere positiva o negativa.
- *Surprise*: una sorpresa più intensa, che porta a un apprezzamento e a un cambiamento di prospettiva.
- *Amazement*: una forma estrema di sorpresa, che può causare disorientamento.

tamento e una forte reazione fisica ed emotiva.

- **Sadness (Tristezza):**

- *Pensiveness*: una forma lieve di tristezza, che può derivare da aspettative non soddisfatte.
- *Sadness*: una sensazione di abbattimento o malinconia che può interferire con il benessere quotidiano.
- *Grief*: una tristezza profonda, spesso associata a una perdita significativa, che porta a un dolore emotivo duraturo.

- **Disgust (Disgusto):**

- *Boredom*: una forma lieve di disgusto, che si manifesta come una repulsione minima a qualcosa di sgradevole.
- *Disgust*: una sensazione forte di repulsione che può causare il desiderio di allontanarsi da un oggetto o una situazione.
- *Loathing*: una forma intensa di disgusto che può provocare reazioni fisiche.

- **Anger (Rabbia):**

- *Annoyance*: una forma lieve di rabbia, che si verifica quando gli obiettivi sono ostacolati in modo non grave.
- *Anger*: una reazione emotiva più forte che porta al desiderio di esprimere disapprovazione o di cambiare la situazione.
- *Rage*: una rabbia intensa, incontrollabile e spesso distruttiva, che può portare a esplosioni di violenza o aggressività.

- **Anticipation (Anticipazione):**

- *Interest*: una forma lieve di anticipazione, in cui una persona è curiosa riguardo a un evento futuro.
- *Anticipation*: una forma più intensa di anticipazione, in cui l'incertezza

o il dubbio dominano l'emozione.

- *Vigilance*: una forma intensa di anticipazione, caratterizzata da un'elevata energia e da un forte desiderio di sperare qualcosa di positivo.

Stati Affettivi

Oltre alle emozioni primarie e alle loro variazioni, Plutchik introduce nel suo modello anche alcuni stati affettivi composti, non propriamente emotivi (che infatti collocava all'esterno dei petali delle emozioni), correlati a due emozioni (i.e., petali) vicine nella ruota. Questi includono:

- **Love** = Joy + Trust
- **Submission** = Fear + Trust
- **Awe** = Fear + Surprise
- **Disapproval** = Surprise + Sadness
- **Remorse** = Sadness + Disgust
- **Contempt** = Disgust + Anger
- **Aggressiveness** = Anger + Anticipation
- **Optimism** = Anticipation + Joy

Abbiamo considerato questa ricchezza semantica del modello di Plutchick adatta più di altri alla rappresentazione delle variazioni emotive spesso sottili nel fenomeno del grooming online.

2.5 Generazione dei Data Set

Nel corso della fase sperimentale sono stati utilizzati due data set, generati sinteticamente attraverso paradigmi di In-Context Learning (ICL) [113] e prompting [114], sfruttando le capacità few-shot (i.e., imparando da pochi esempi) di modelli linguistici avanzati come *GPT-3* (OpenAI) [61] e *LLaMA 2 uncensored (Meta)* [145]. Il primo data set associa ad ogni frase, l'etichetta

dell’emozione primaria e secondaria corrispondente o lo stato affettivo, dove presente, seguendo lo schema di Plutchik [40]. Questo, è stato suddiviso in due sottoinsiemi distinti prima dell’avvio di qualsiasi fase di addestramento. La prima porzione (parte 1), più consistente, è stata utilizzata per le attività di *fine-tuning* (v. sez. 3.2) necessarie all’addestramento delle diverse componenti del modello. La seconda porzione (parte 2), completamente indipendente e mai osservata dal modello durante l’intero processo di addestramento, è stata mantenuta separata al fine di effettuare valutazioni successive sull’architettura finale. In particolare, tale insieme è stato impiegato per la valutazione del rilevamento degli stati affettivi composti che, dipendendo in parte dalle predizioni fornite dalle teste di classificazione sulle emozioni primarie e secondarie, richiedevano dati completamente nuovi e non contaminati, così da evitare ogni forma di bias indotto [146] dal riutilizzo di dati già elaborati durante la fase di sviluppo e garantire una stima più realistica delle prestazioni generali del sistema. Il secondo data set [147], invece, già presente allo stato dell’arte e relativo alle fasi del grooming, è stato usato per l’estrazione di relazioni tra i pattern emozionali e le fasi/tag del processo di adescamento (v. sez. 1.1). La generazione dei dati ha beneficiato delle caratteristiche complementari dei due modelli: *GPT-3* ha prodotto testi più formali e generalisti, mentre *LLaMA 2 uncensored* ha fornito output più esplicativi e diretti, rivelatisi fondamentali per rappresentare fedelmente le dinamiche emozionali e linguistiche dell’ambito studiato.

I prompt utilizzati per la generazione sono stati strutturati e contestualizzati, includendo esempi in linea con le etichette emozionali target e la sintassi desiderata. Questo approccio ha permesso di sfruttare la conoscenza già presente nei modelli, adattandola al task specifico (v. sez. 3.3).

I data set sono stati salvati in formato Comma-Separated Value (CSV) e gestiti in ambiente Python tramite la libreria *pandas* [148].

Le caratteristiche comuni ai data set includono:

- **Etichettatura semantica con verifica**, supportata da controlli post-generativi automatici e manuali basati su analisi vettoriali e congruenza testo-etichetta.
- **Eterogeneità linguistica e copertura contestuale**, con diversificazione stilistica, presenza di strutture complesse e fenomeni linguistici critici (negazioni, ambiguità, impliciti, linguaggio inappropriato).
- **Pulizia e qualità testuale**, ottenute tramite rimozione di duplicati, verifica sintattica, filtraggio di incoerenze e outlier.

Data set 1: final_emotions_dataset

Il data set generato contiene 26.631 esempi ed è strutturato in tre colonne:

- **sentence**: contiene frasi in linguaggio naturale, varie e diversificate per contenuto e stile.
- **emotion_label1**: include le otto emozioni primarie di Plutchik, arricchite dalle diverse intensità e dalla classe *neutral*.
- **emotion_label2**: fornisce una mappatura di *emotion_label1* nelle corrispondenti emozioni primarie, aggregando intensità e varianti in classi primarie.

Facts

Di seguito si propone un'analisi sia quantitativa che qualitativa di questo data set.

Distribuzione quantitativa

La distribuzione delle etichette nel data set è rappresentata tramite radar graph in Fig. 2.2, 2.3, 2.4; dal confronto tra i grafici emerge chiaramente come il numero di campioni associati alle etichette primarie (Figura 2.2) è significativamente più elevato rispetto a quello delle altre etichette (Figure 2.3

e 2.4). Questa maggiore rappresentatività garantisce un numero sufficiente di esempi per l’addestramento efficace della testa di classificazione responsabile dell’identificazione dell’emozione primaria (v. componente num. 1 in fig. 3.3) e parte centrale dell’architettura finale.

In entrambe le rappresentazioni (Figure 2.2 e 2.3), l’etichetta *neutral* si distingue per una distribuzione non uniforme rispetto alle altre.

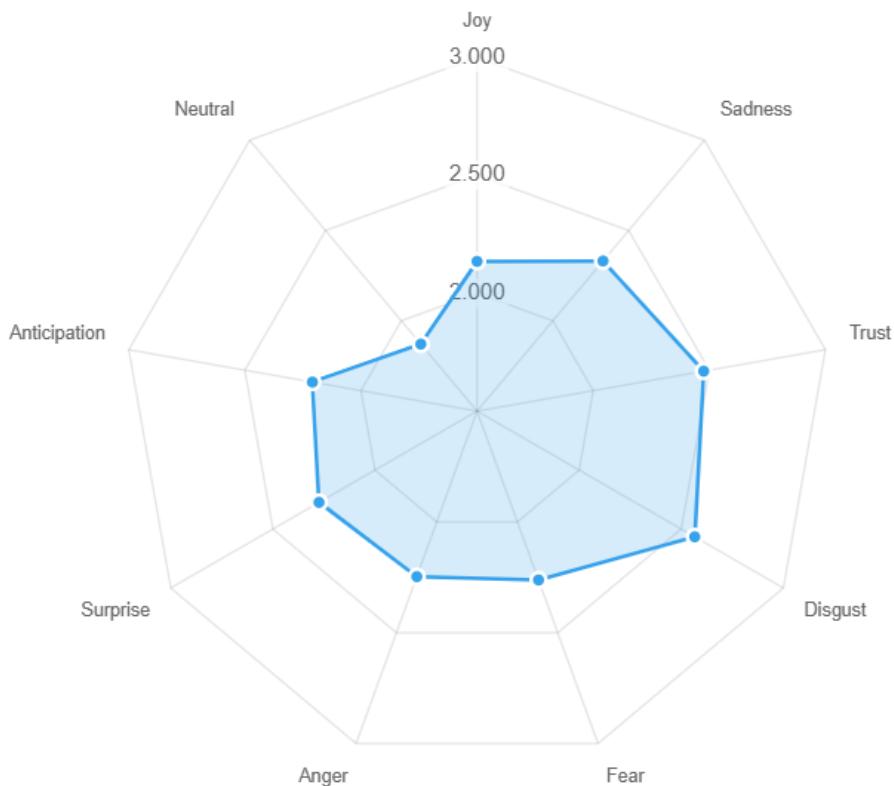


Figura 2.2: Distribuzione delle etichette in *emotion_label2* (emozioni primarie).

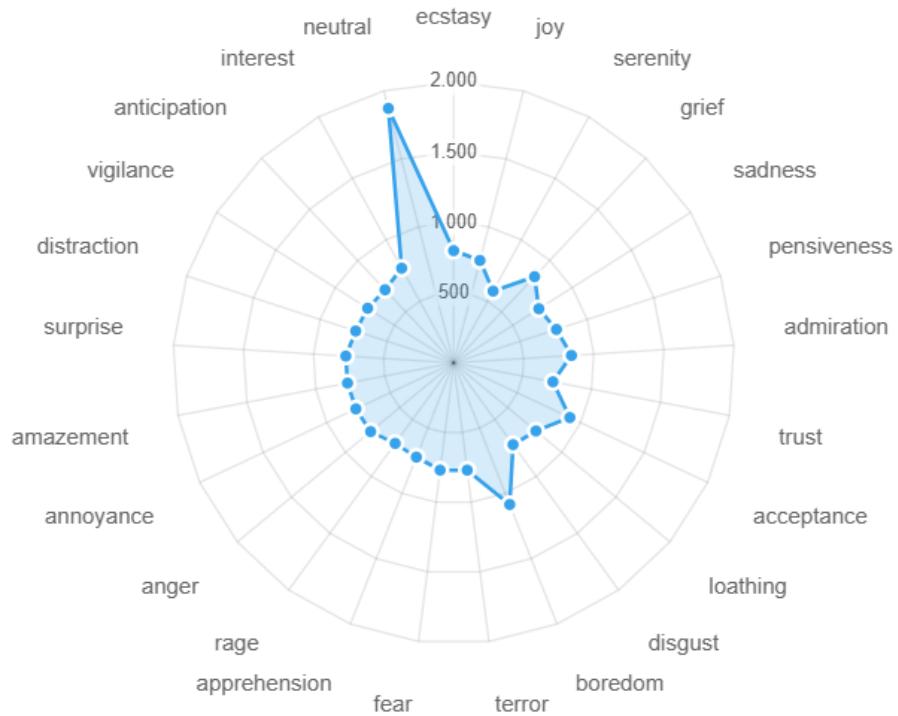


Figura 2.3: Distribuzione delle etichette in *emotion_label1* (emozioni secondarie).

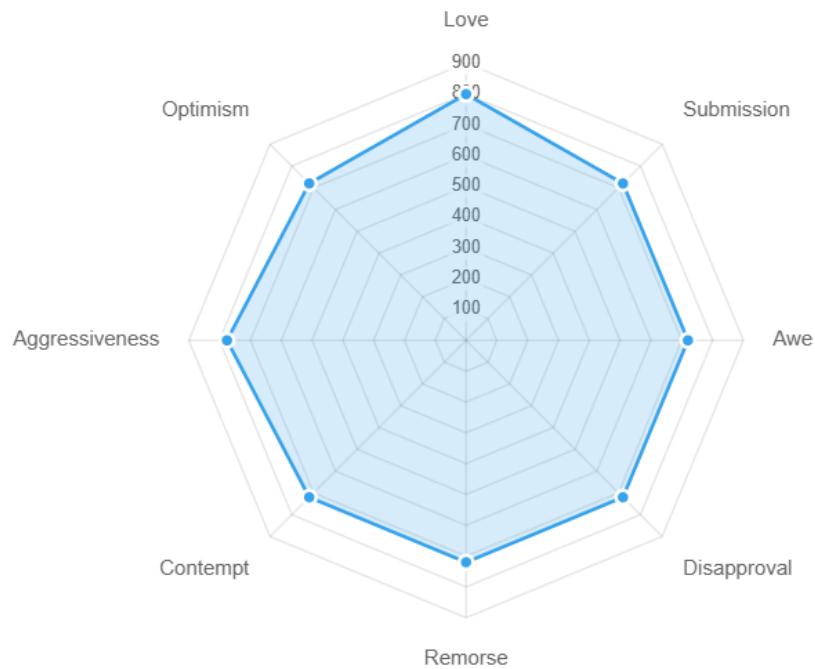


Figura 2.4: Distribuzione delle etichette in *emotion_label1* (stati emotivi).

La tabella 2.4 presenta le statistiche descrittive del data set, con particolare attenzione ai parametri relativi alla lunghezza delle frasi, espressi sia in

numero di parole che in numero di caratteri e alla loro rappresentatività linguistica. Tali dati offrono una panoramica quantitativa della distribuzione delle lunghezze, evidenziando la variabilità e la struttura del corpus analizzato.

Lunghezza delle frasi (in parole)		Lunghezza delle frasi (in caratteri)	
Lunghezza media	6,98	Lunghezza media	37,64
Deviazione standard	2,00	Deviazione standard	11,17
Lunghezza minima	1	Lunghezza minima	4
1° quartile (25%)	6	1° quartile (25%)	29
Mediana (50%)	7	Mediana (50%)	36
3° quartile (75%)	8	3° quartile (75%)	43
Lunghezza massima	28	Lunghezza massima	140

Tabella 2.4: Statistiche descrittive del data set *final_emotions_dataset* sulle lunghezze delle frasi.

Abbiamo scelto di generare frasi brevi come item del data set, per facilitare l’addestramento mirato delle emozioni, riducendo la complessità sintattica e minimizzando possibile rumore. In particolare, la maggior parte delle frasi presenta una struttura concisa: la mediana della lunghezza delle frasi è pari a 7 parole e 36 caratteri, indicando che il 50% delle frasi non supera tali soglie. Inoltre, il primo quartile (25%) corrisponde a 6 parole e 29 caratteri, mostrando che un quarto del corpus è costituito da frasi ancora più brevi, con una distribuzione centrata su lunghezze ridotte.

Analisi qualitativa

Abbiamo, inoltre, condotto un’analisi automatizzata per valutare la congruenza tra il contenuto semantico delle frasi e le etichette emozionali ad esse associate durante la generazione del data set, applicando un Sentence Embedding con il modello preaddestrato *sentence-t5-large* [123], che consente di ottenere rappresentazioni vettoriali dense (embedding [56]) delle frasi

nello spazio semantico latente. Il processo include i seguenti passaggi:

1. Encoding delle frasi in rappresentazioni dense tramite sentence-t5-large.
2. Encoding delle etichette emozionali, utilizzando le loro denominazioni testuali (e.g., *paura*, *gioia*, *disgusto*) rappresentate nello stesso spazio embedding.
3. Calcolo della *cosine similarity*, metrica già descritta nella sezione 2.1, tra ogni frase e la rispettiva etichetta, per stimare il grado di affinità semantica.

La distribuzione degli embedding è stata analizzata mediante riduzione dimensionale con PCA (Principal Component Analysis), come descritto nell'articolo “Efficient tools for principal component analysis of complex data” [149], per investigare la struttura nello spazio. Le figure illustrate mostrano la *similarità coseno* media per emozioni primarie, stati affettivi composti e secondarie (Figure 2.5, 2.6, 2.7). I valori di similarità, compresi tra 0.72 e 0.79 per le emozioni primarie, tra 0.734 e 0.799 per gli stati affettivi composti e tra 0.720 e 0.789 per quelle secondarie, indicano una robusta coerenza semantica tra le frasi nel data set completo e le labels emotive associate (v sez. 2.4).

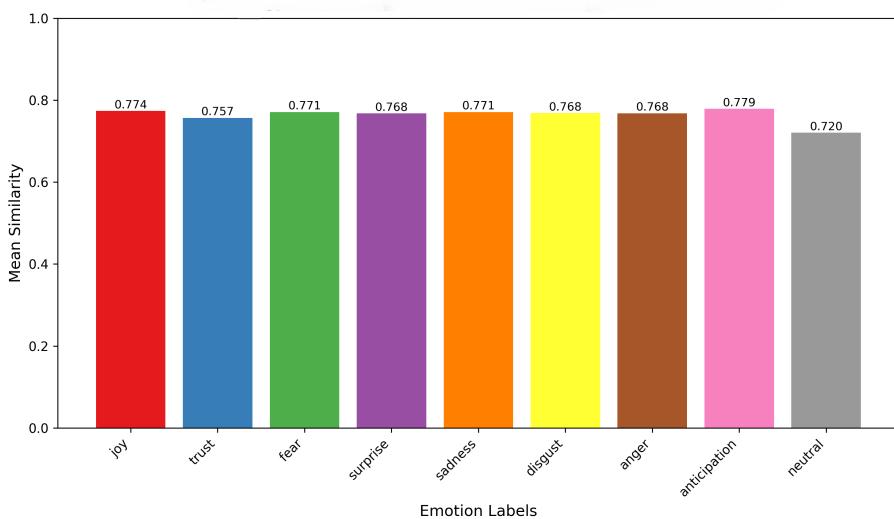


Figura 2.5: Similarità semantica emozioni primarie con Sentence T5.

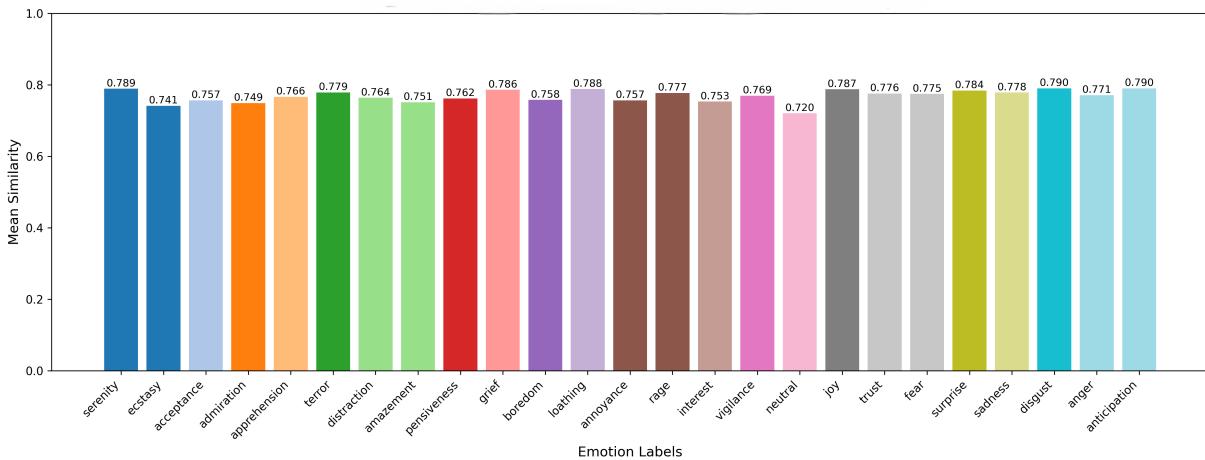


Figura 2.6: Similarità semantica emozioni secondarie con modello Sentence T5.

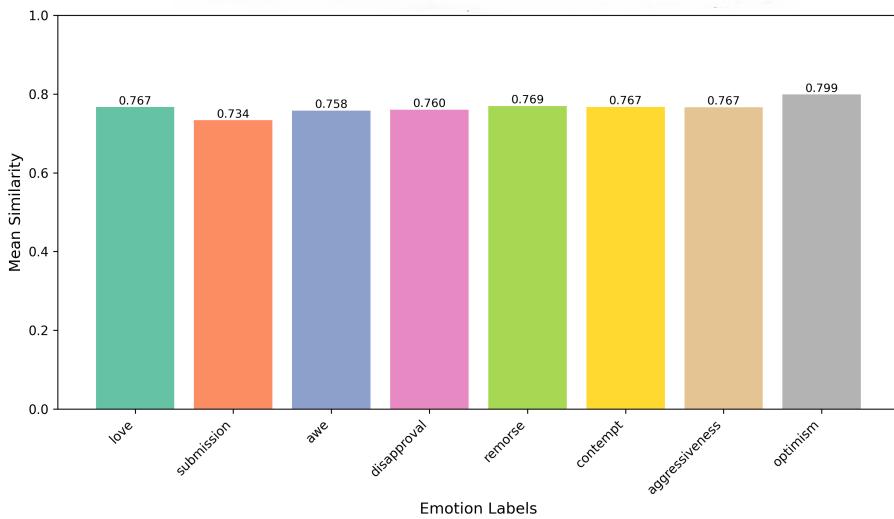


Figura 2.7: Similarità semantica stati affettivi composti con modello Sentence T5.

Parte 1

Questa sezione del data set, è la risorsa principale utilizzata per l'addestramento del modello gerarchico ed è composta da 23.331 samples (87,5% del totale).

Distribuzione delle emozioni nella parte 1

Le classi usate per training e test sono bilanciate, mantenendo inalterata la

distribuzione originale. La tabella 2.5 mostra le frequenze delle etichette in *emotion_label1* e *emotion_label2* relativamente a questa parte di data set, evidenziandone l'organizzazione gerarchica, con l'indipendenza degli stati affettivi e della classe *neutral*.

Tabella 2.5: Suddivisione delle etichette nella Parte 1 del data set

Emozione	Frequenza (emotion_label1)	Frequenza (emotion_label2)
ecstasy	699	—
joy	653	1835
serenity	483	—
grief	742	—
sadness	619	2031
pensiveness	670	—
admiration	741	—
trust	620	2176
acceptance	815	—
loathing	661	—
disgust	620	2266
boredom	985	—
terror	670	—
fear	670	1962
apprehension	622	—
rage	609	—
anger	668	1947
annoyance	670	—
amazement	670	—
surprise	670	1974
distraction	634	—
vigilance	626	—

Continua a pagina successiva...

Tabella 2.5 — continuazione

anticipation	614	1909
interest	669	—
aggressiveness	676	676
love	699	699
contempt	620	620
disapproval	620	620
awe	620	620
optimism	620	620
submission	620	620
remorse	620	620
neutral	1771	1771

Parte 2

Questa sezione del data set principale rappresenta il 12,5% (3300 esempi) del totale ed è stata estratta prima dell'addestramento, esclusivamente per la validazione finale del modello su frasi mai viste prima e la valutazione della logica di rilevamento degli stati affettivi.

Distribuzione delle emozioni nella parte 2

Nella tabella 2.6 si riporta la distribuzione delle labels in questa sottosezione di data set.

Tabella 2.6: Suddivisione delle etichette nella Parte 2 del data set

Emozione	Frequenza (emotion_label1)	Frequenza (emotion_label2)
ecstasy	100	—
joy	100	300
serenity	100	—
grief	100	—

Continua a pagina successiva...

Tabella 2.6 — continuazione

sadness	100	300
pensiveness	100	—
admiration	100	—
trust	100	300
acceptance	100	—
loathing	100	—
disgust	100	300
boredom	100	—
terror	100	—
fear	100	300
apprehension	100	—
rage	100	—
anger	100	300
annoyance	100	—
amazement	100	—
surprise	100	300
distraction	100	—
vigilance	100	—
anticipation	100	300
interest	100	—
aggressiveness	100	100
love	100	100
contempt	100	100
disapproval	100	100
awe	100	100
optimism	100	100
submission	100	100
remorse	100	100

Continua a pagina successiva...

Tabella 2.6 — continuazione

neutral	100	100
---------	-----	-----

Il bilanciamento del data set è una sfida comune nel fine-tuning [105], soprattutto quando le classi del data set non sono equamente distribuite. Un data set sbilanciato può portare a un modello che tende a favorire la classe dominante, ignorando segnali significativi appartenenti alla classe minoritaria. Per affrontare questo problema, esistono diverse tecniche, come la "pesatura delle classi", che assegna un peso maggiore alle classi meno rappresentate, incentivando il modello a prestare maggiore attenzione a questi esempi [150]. Altre strategie includono l'oversampling, che aumenta il numero di esempi appartenenti alle classi minoritarie, e l'undersampling, che riduce il numero di esempi appartenenti alle classi maggioritarie per ottenere un data set bilanciato [151]. Per garantire questo bilanciamento, in tutti gli esperimenti è stato utilizzato un approccio di tipo undersampling.

Data set 2: grooming_sintetic_emotions_data set

Il *grooming_sintetic_emotion_data set* [19] è composto da 2.580 samples ed è strutturato in quattro colonne principali: id, sentence, tag e phase. Costituisce il data set di test a cui è stato applicato il modello finale ed è stato sviluppato tramite *Llama-3.1-8B-instruct-abliterated* [152] con In-context Learning [113] e tecniche di adattamento sequenziale dei Prompt [114]. L'intero processo di generazione e analisi del data set, è riportato nello studio di Ludovico Guercio [147], dal quale è stato estrappolato, al fine di applicarvi il modello gerarchico di classificazione sviluppato in questa tesi e analizzarne i pattern emozionali derivanti.

Facts

Le colonne del data set sono così definite:

- **id**: un identificatore univoco per ciascun esempio del data set, utile per tracciare ogni singolo dato.
- **sentence**: la frase che rappresenta il dialogo tra due interlocutori, in cui "G" è la persona che sta cercando di instaurare una relazione di grooming, mentre "V" è la vittima, che risponde alle sollecitazioni del predatore.
- **tag**: una categoria che descrive l'attività svolta durante la conversazione.
- **phase**: indica a quale fase del processo di grooming appartiene la conversazione.

Come visto nel capitolo [1], gli adescatori operano tipicamente seguendo un processo strutturato in sei fasi principali (*phases*, v. sez. 1.1), che rappresentano le tappe fondamentali della strategia di manipolazione e adescamento adottata. Tale processo sequeziale è caratterizzato da strategie mirate a guadagnare la fiducia, creare dipendenza, isolare la vittima e infine esercitare controllo e abuso. Sono definiti otto comportamenti specifici in questo data set, con l'intestazione di *tag*, che rappresentano le principali attività del groomer per ciascuna *fase*, derivate dal data set di Perverted Justice [153], alla base della generazione degli esempi sintetici del data set.

L'analisi di questi elementi consente di comprendere i meccanismi della manipolazione e di individuare segnali di pericolo.

1. **Activities** (Phase 1 – Targeting and Trust Building)

Il predatore raccoglie informazioni sulla routine della vittima (orari scolastici, impegni familiari, hobby) per identificare momenti di vulnerabilità e opportunità di contatto.

2. **Personal Information** (Phase 1 – Targeting and Trust Building)

Il groomer cerca di ottenere dati sensibili, come indirizzo, numero di telefono o informazioni sulla famiglia, utili per stabilire un contatto diretto o per future minacce.

3. **Compliments** (Phase 2 – Fulfilment of a Need)

L'uso di lodi e attenzioni, spesso esagerate, rafforza il legame emotivo con la vittima e abbassa le sue difese psicologiche, facendola sentire speciale e apprezzata.

4. Relationship (Phase 1 and 2 – Targeting and Trust Building, Fulfillment of a Need)

Il predatore costruisce un rapporto di fiducia fingendosi amico o possibile partner, ponendo domande sulle relazioni affettive della vittima per comprenderne la vulnerabilità emotiva.

5. Reframing (Phase 3 – Isolation)

Il discorso viene progressivamente orientato verso contenuti intimi o sessuali, con passaggi graduali e apparentemente naturali, per superare le resistenze della vittima.

6. Desensitization (Phase 3 – Isolation)

La vittima viene esposta in modo progressivo a contenuti sessuali o intimi, al fine di normalizzare questi argomenti e ridurre il suo senso di disagio iniziale.

7. Isolation (Phase 3 and 5 – Isolation, Control)

Il groomer allontana la vittima dal contesto familiare e sociale, minimizzando l'importanza di figure adulte di riferimento e rafforzando la dipendenza emotiva.

8. Approach (Phase 4 and 5 – Sexualisation of the Relationship and Abuse, Control)

Il predatore avanza richieste sessualmente esplicite, sollecita immagini intime o propone incontri fisici con l'obiettivo di concretizzare l'abuso.

Nel testo, di cui si propone un estratto di seguito, ogni messaggio ha un *tag* e le relative *fasi* nel quale si colloca (v sez. 1.1); un messaggio con un *tag* può comparire anche in più *fasi*.

```

23,"G: so u like that color bcs it's very pretty ❤️ | V: no, just because I like it",reframing,isolation
24,"G: aw man, ur so sweet 😊 | V: stop saying that 😊",reframing,isolation
25,"G: u r not like the other girls, you're special ❤️ | V: what are you talking about?",reframing,isolation
26,G: because you're so mature and beautiful 🌟 | V: stop it! 😊,reframing,isolation
27,G: i want to take care of u like a mom does 😊 | V: that's weird,reframing,isolation
28,G: but I'm sure you'll feel much better with me ❤️ | V: this is too much for me 😊,reframing,isolation
29,G: u know I'll always be there 4u | V: stop saying that 🌟,reframing,isolation
30,G: we're best friends now ❤️ | V: you're just being weird 😊,reframing,isolation
31,"G: don't tell ur mom about it, keep it a secret 😊 | V: why do I need to hide something?",reframing,isolation
32,G: because I know u better than they do ❤️ | V: that's not true 😊,reframing,isolation
33,G: ur so brave and I love it ❤️ | V: stop saying those things,reframing,isolation

```

Figura 2.8: Estratto data set sintetico chat di grooming

Questo data set risulta fondamentale per l'applicazione del sistema di classificazione implementato a conversazioni sintetiche realistiche, permettendo un'analisi accurata della correlazione tra emozioni e fasi/tag di adescamento. Di seguito, nelle tabelle 2.7 il numero di esempi per ciascun valore delle colonne tag e phase (fase).

Frequenze delle Fasi		Frequenze dei Tag
targeting_and_gaining_trust	1003	isolation 365
isolation	917	communicative_desensitization 358
fulfilling_needs	351	personal_information 357
sexualizing_relationship	308	compliment 351
		relationship 330
		activities 316
		approach 308
		reframing 194

Tabella 2.7: Frequenze delle fasi e dei tag nel dataset

Capitolo 3

Esperimenti

Il processo sperimentale è stato articolato in fasi sequenziali, come illustrato in Figura 3.1, dove vediamo rappresentato il workflow con i diversi step di sviluppo che hanno condotto alla definizione dell'architettura finale del sistema, mostrata in Figura 3.3. Ciascuna fase sperimentale sarà descritta e analizzata in dettaglio nelle sezioni successive.

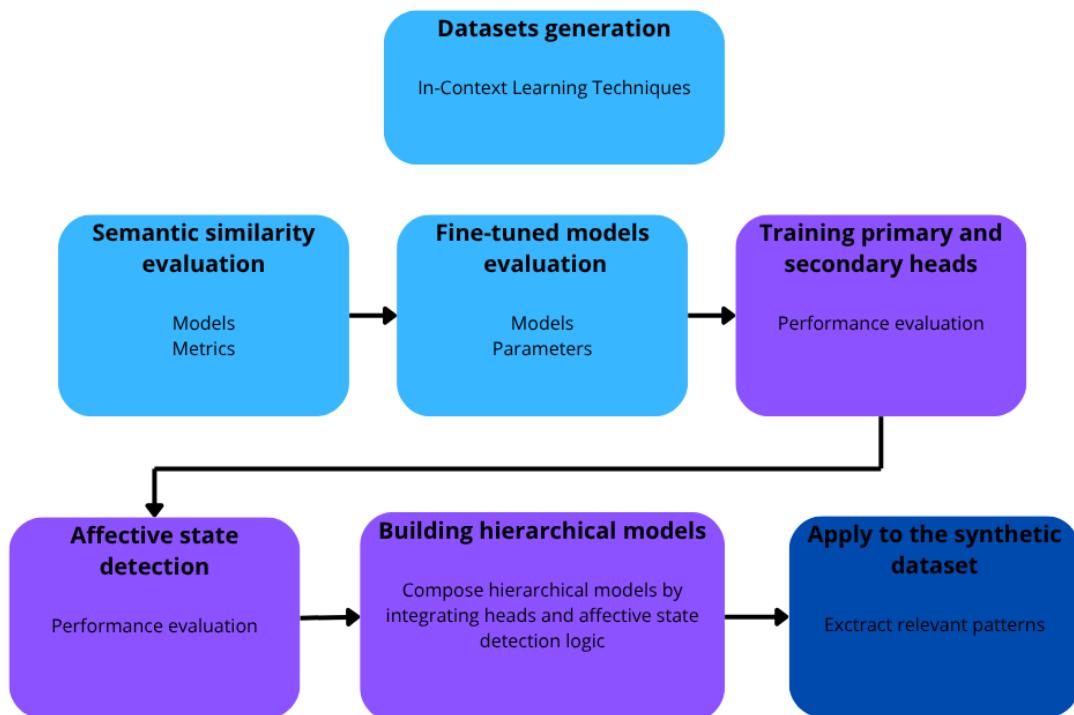


Figura 3.1: Flusso degli esperimenti

L'approccio seguito è di tipo bottom-up: si parte da configurazioni di base per poi affinare gradualmente le soluzioni, attraverso l'integrazione di modelli e tecniche più sofisticati. Questa strategia incrementale consente un ablation study [154], ossia l'individuazione del contributo di ciascuna dimensione di analisi al risultato finale. Il sistema finale si compone di tre macro-aree, ciascuna con un ruolo specifico nel processo di classificazione emozionale, numerate in figura 3.3. La prima componente del sistema è responsabile dell'assegnazione dell'etichetta corrispondente all'emozione primaria. La seconda macro-area è dedicata alla classificazione dell'emozione secondaria, mentre la terza componente si occupa del rilevamento di stati affettivi complessi (v. sez. 2.4), nel caso in cui siano presenti due emozioni tra loro affini.

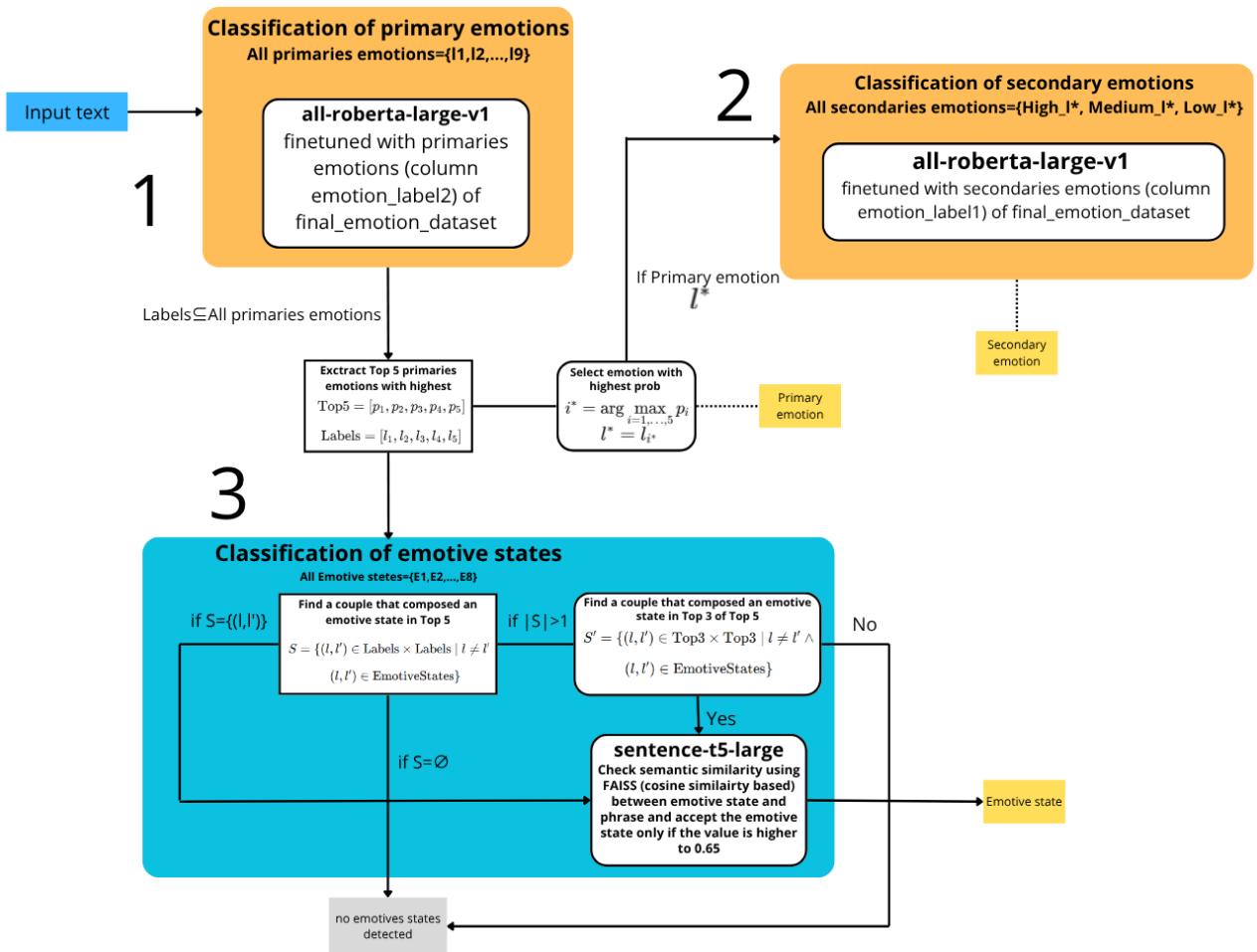


Figura 3.2: Flusso di esecuzione nell'architettura sviluppata.

La Tabella 3.1 fornisce una panoramica sintetica dei principali esperimenti condotti, riportando per ciascuno di essi il data set utilizzato, l’obiettivo perseguito e i modelli usati.

Tabella 3.1: Dettagli generali degli esperimenti condotti.

Esperimento	data set	Obiettivo	Modelli
1	final_emotions_dataset (parte 1)	Valutazione modelli e metriche migliori semantic similarity	BERT, Sentence-Transformers
2	final_emotions_dataset (parte 1)	Valutazione modelli e metriche migliori fine-tuning	BERT, Sentence-Transformers
3	final_emotions_dataset (parte 1 e 2)	Classificazione emozioni primarie e secondarie	RoBERTa
4	final_emotions_dataset (parte 1 e parte 2)	Classificazione stati emotivi	Sentence-T5
5	grooming_sintetic_emotions_dataset	Applicazione classificatore gerarchico completo ai dati sintetici	Architettura gerarchica implementata

Ogni esperimento rappresenta uno step fondamentale nel processo di sviluppo, in un’ottica gerarchica di progressivo dettaglio.

3.1 Scelta Modelli e Metriche per la Prossimità Semantica

Dopo lo studio preliminare presentato nella sezione 2.1, che ha permesso di valutare l’efficacia delle diverse metriche relativamente alla similarità semantica e restringere il campo di ricerca, in questo esperimento abbiamo testato l’algoritmo K-Nearest Neighbors (KNN) [90], con tre diverse misure di similarità nello spazio vettoriale: *cosine similarity* [75], *FAISS inner product* [85] e *Roger-Tanimoto similarity* [82] (v. sezione 2.3).

Con una prima fase sperimentale scegliamo la misura migliore per la similarità semantica [44] nel nostro contesto, calcolata a livello di embedding vettoriali [56], per la classificazione automatica delle emozioni basata sul modello di Plutchik (v. sez. 2.4). In particolare, si mira a determinare quale modello linguistico consenta di riconoscere con maggiore accuratezza la similarità semantica tra frasi, al fine di utilizzarlo successivamente per misurare la prossimità tra una frase e un’emozione specifica.

L’esperimento valuta l’accuratezza della classificazione per diversi valori di k . Un’accuratezza elevata indica che gli embedding del modello sono in grado

di raggruppare frasi semanticamente simili nello spazio vettoriale, riflettendo con precisione le emozioni secondo il modello di Plutchik [40], fornendo una base solida per l’ottimizzazione della classificazione emozionale nei successivi esperimenti; la tabella 3.2 riporta le configurazioni dell’esperimento.

Tabella 3.2: Configurazioni per la valutazione della prossimità semantica

Parametro	Valore
data set	<code>final_emotion_data_set</code> (parte 1)
Numero di Classi	9 emozioni (<i>anger, anticipation, disgust, fear, joy, neutral, sadness, surprise, trust</i>)
Campionamento	500 frasi per emozione
Numero di esempi	4.500
Modelli utilizzati	bert-base-uncased, all-distilroberta-v1, distiluse-base-multilingual, paraphrase-mpnet-base-v2, paraphrase-ml-mpnet-base-v2, all-roberta-large-v1, all-MiniLM-L6-v2, all-MPNet-base-v2, paraphrase-distilroberta-base-v1, multi-qa-mpnet-base-dot-v1, stsb-roberta-large, sentence-t5-large
Tokenizer	BertTokenizer (BERT), BertTokenizer ottimizzato (SBERT)
Max Length	512 (BERT), 256 (SBERT)
Batch Size	Singola frase
Librerie	numpy, pandas, sentence-transformers, transformers, torch, sklearn, matplotlib, seaborn
Hardware	CPU Intel(R) Core(TM) i7-1165G7, RAM 8 GB
Valori di k	1, 2, 3, 4, 5
Metriche	Cosine Similarity, FAISS Inner Product, Roger-Tanimoto Similarity

Workflow

L’esperimento valuta le prestazioni di 12 modelli di embedding nella classificazione delle 8 emozioni primarie (*anger, anticipation, disgust, fear, joy, sadness, surprise, trust*) e della label *neutral*, basate sulla colonna *emotion_label2* del data set, al fine di valutare quale modello riesce a garantire una migliore analisi della similarità semantica. Inizialmente, il data set (v. sez. 2.5) viene caricato e filtrato per includere solo le frasi con etichette appartenenti alle 9 categorie emozionali. Per bilanciare le classi, sono state campionate 500 frasi per ciascuna emozione, per un totale di 4.500 frasi. Le frasi sono codificate in vettori densi utilizzando 12 modelli di embedding,

riportati nella tabella 3.2. Per bert-base-uncased [60], gli embedding sono estratti dal pooler output [131] o dal primo token dell’ultimo strato nascosto [132], mentre per gli altri modelli si utilizza la libreria *sentence-transformers* [89], per rappresentazioni ottimizzate per la similarità semantica. La classificazione è eseguita tramite l’algoritmo K-Nearest Neighbors (KNN) [90], testando valori di k nel range [1-5]. Abbiamo quindi confrontato tre metriche di similarità: *cosine similarity*, basata sul prodotto scalare normalizzato tra vettori; *FAISS inner product*, implementata tramite la libreria FAISS per ricerche efficienti su vettori normalizzati e la misura di similarità di *Roger-Tanimoto*, che adatta la distanza di Tanimoto a vettori continui tramite binarizzazione basata sulla mediana (v. sez. 2.1). Per ogni frase, si identificano i k vicini più simili nel set di training e l’etichetta è assegnata in base alla moda delle etichette dei vicini (v. dettaglio nella sezione 2.3). Le prestazioni sono valutate calcolando l’*accuracy* per ciascun valore di k e generando *matrici di confusione* (v. sez. 2.2) per ogni modello e metrica. I dettagli sui parametri, inclusi i valori di k e le configurazioni dei modelli, sono riportati nella tabella 3.2. nella figura 3.3 si propone una vista aggregata del flusso di esecuzione appena descritto.

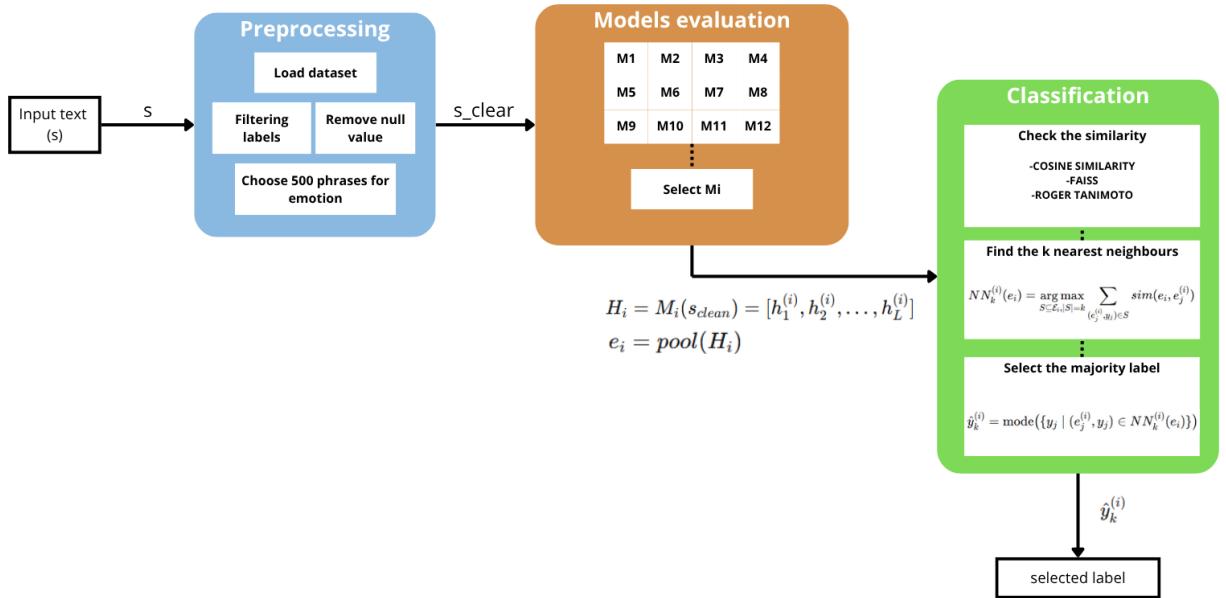


Figura 3.3: Flusso di esecuzione dell’esperimento 1.

3.2 Scelta modelli e parametri per fine-tuning

L’obiettivo principale di questo confronto è valutare le prestazioni di modelli basati su architetture transformer e sentence-transformer (v. sez. 2.3 per la classificazione automatica delle emozioni, riconosciute all’interno del linguaggio naturale, attraverso l’applicazione della tecnica del fine-tuning e del transfer-learning (v. sec. 2.2). L’esperimento si basa sull’utilizzo del data set etichettato precedentemente descritto (vedi sezione 2.5). Il focus della ricerca è l’adattamento di modelli pre-addestrati [124] al compito specifico di classificazione di emozioni e stati affettivi composti, che verranno poi applicati al riconoscimento delle fasi del grooming [69].

Nell’esperimento, sono state definite alcune caratteristiche generali, sintetizzate nella tabella 3.3, che costituiscono il setup di base utilizzato per l’analisi dei vari modelli.

Tabella 3.3: Configurazioni complete per l’Esperimento 2

Parametro	Valore
data set	final_emotion_data set.csv (parte 1)
Numero di Classi	9 emozioni (Joy, Sadness, Fear, Anger, Surprise, Disgust, Trust, Anticipation, Neutral)
Campioni per Etichetta	400
Numero di Esempi	3.600
Epoche di Addestramento	5, 7
Learning rate	2.000×10^{-5} , 3.000×10^{-5}
Weight decay	0.01, 0.001
Modelli Utilizzati	AutoModelForSequenceClassification con: paraphrase-mpnet, multilingual-paraphrase-mpnet, RoBERTa, distilbert, bert-base-uncased, all-MiniLM-L6-v2, all-MPNet-base-v2, paraphrase-distilroberta-base-v1, multi-qa-mpnet-base-dot-v1, stsbert-large, distiluse-base-multilingual-cased, sentence-t5-large
Tokenizer	AutoTokenizer associato al modello
Massima Lunghezza Sequenze	256
Batch Size	16, 32 (ricerca iperparametrica)
Tecniche	Fine-tuning supervisionato con early stopping; bilanciamento delle classi per intensità emozionale
Metriche	Accuracy, F1-score (weighted), Precision (weighted), Recall (weighted), Matrice di confusione
Librerie	numpy, pandas, sentence-transformers, transformers, torch, scikit-learn, matplotlib, seaborn, data set
Hardware	CPU Intel(R) Core(TM) i7-1165G7, RAM 8 GB
Framework di Addestramento	Hugging Face Transformers Trainer

Workflow

Il fine-tuning è stato condotto sui modelli pre-addestrati riportati nella Tabella 3.3, ottimizzando la testa specifica per il compito di classificazione delle emozioni. Il workflow sperimentale si articola nelle seguenti fasi:

- **Preprocessing:** caricamento del data set, rimozione di duplicati e valori nulli, normalizzazione (i.e., riduzione a forma standardizzata) del testo.
- **Bilanciamento:** campionamento uniforme per ciascuna combinazione di classe emozionale e livello di intensità.
- **Codifica:** trasformazione delle etichette testuali in valori numerici mediante *LabelEncoder* [155].
- **Suddivisione del data set:** divisione stratificata in set di training (70%), validazione (10%) e test (20%).
- **Ottimizzazione iperparametrica:** esplorazione delle combinazioni di learning rate (2.000×10^{-5} , 3.000×10^{-5}) [156], batch size (16, 32) [157], weight decay (0.01, 0.001) [158] e numero di epoche (5, 7) [159].
- **Fine-tuning:** addestramento della sola testa classificatrice, mantenendo congelati i pesi del backbone.
- **Early stopping:** interruzione anticipata dell'addestramento in assenza di miglioramento sulla loss di validazione.
- **Valutazione:** analisi tramite accuracy, F1-score, precision e recall (tutte in versione weighted), con il supporto di matrici di confusione e curve di apprendimento (vedi sez. 2.2).

L'addestramento è stato eseguito utilizzando le librerie *transformers* [160], *torch* [161] e *datasets* [162], mentre l'analisi dei risultati è stata supportata da *scikit-learn* [163], *pandas* [148], *numpy* [164], *matplotlib* [165] e *seaborn* [166]. L'esperimento si concentra sulla classificazione delle otto emozioni primarie, con l'aggiunta della classe *Neutral*. Per ciascuna emozione sono stati selezionati 400 esempi per ognuno dei tre livelli di intensità previsti dal

modello di Plutchik, ottenendo 1.200 frasi per emozione e un totale di 10.800 istanze. Avendo identificato i modelli e gli iperparametri più performanti, per gli esperimenti successivi ci siamo concentrati sul fine-tuning del modello *RoBERTa* [122], risultato il migliore con un netto divario di accuracy rispetto a tutti gli altri. I risultati specifici, sono mostrati nella sezione 4.2.

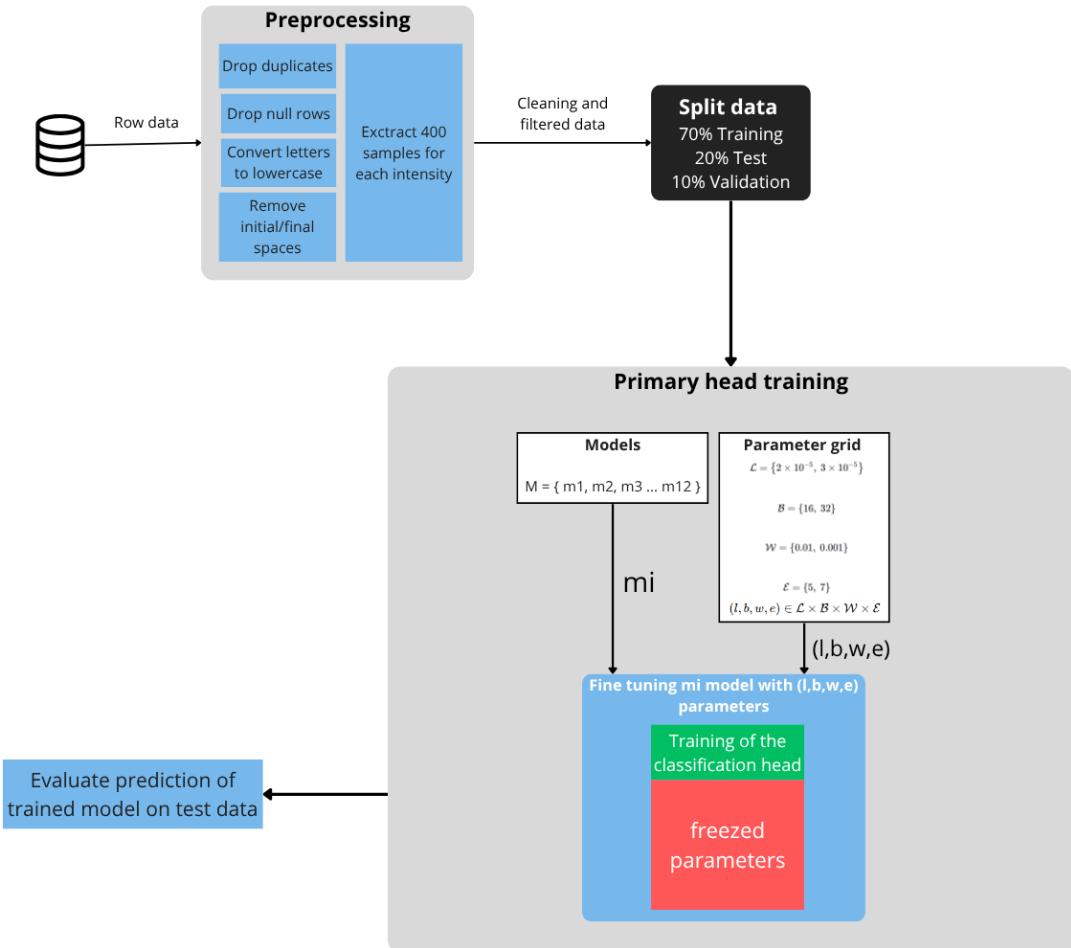


Figura 3.4: Diagramma di flusso dell’addestramento e confronto dei vari modelli

3.3 Classificazione emozioni primarie e secon- darie

Dopo aver confrontato gli strumenti per portare avanti le fasi della classificazione, in questo passo della sperimentazione implementiamo la classificazione gerarchica delle emozioni. Ci basiamo su architetture neurali transformer [33], usate per identificare sia l'emozione primaria espressa in un testo, sia la sua specifica intensità emotiva (i.e., emozione secondaria), come nella struttura del modello di Plutchik, secondo cui ogni emozione primaria può manifestarsi con tre livelli di intensità differenti [72].

La pipeline di classificazione è articolata in una fase iniziale di preprocessing del testo, cui segue la tokenizzazione tramite modelli pre-addestrati; successivamente, si applica il processo di classificazione attraverso le teste gerarchiche.

L'intero sistema è stato ottimizzato e valutato sul corpus etichettato con emozioni di Plutchick (vedi 2.5). Per la valutazione delle performance del modello, sono state adottate le metriche standard di classificazione (v. sez. 2.2). In particolare, l'accuracy fornisce una misura generale della correttezza delle predizioni. Tuttavia, per tenere conto dell'eventuale squilibrio tra le classi, sono state considerate anche le metriche di precision, recall e F1-score (v. sez. 2.2), calcolate in modalità weighted. Questa ponderazione assicura una valutazione più bilanciata delle prestazioni, evitando che classi meglio rappresentate influenzino eccessivamente il risultato complessivo ed è stata scelta come elemento standard del workflow per poter essere applicata anche a data set non bilanciati.

Per un'analisi più approfondita delle misclassification, sono state generate e analizzate le matrici di confusione. In particolare, è stata verificata la presenza di errori di classificazione tra emozioni semanticamente affini (che poi chiameremo metaclassi, v. sez. 3.3) oppure tra quelle adiacenti nel

modello di Plutchik.

Per ciascuna classe, è stato calcolato il tasso di veri positivi rispetto ai falsi positivi [167], consentendo di stimare la robustezza e la separabilità delle predizioni. Inoltre, è stata utilizzata la curva ROC [118] per valutare la capacità del modello di distinguere tra le classi.

Le configurazioni principali per l'addestramento e la valutazione del modello sono riassunte nella Tabella 3.4. Il dettaglio su come tali parametri di configurazione sono stati utilizzati durante l'intero ciclo di sviluppo è spiegato nella sezione 3.3.

Tabella 3.4: Parametri sperimentali comuni agli esperimenti di classificazione delle emozioni

Parametro	Valore
Modello linguistico	all-roberta-large-v1
Tokenizzazione	AutoTokenizer associato al modello
data set	final_emotion_data set (parte 1 e parte 2)
Lunghezza massima sequenze	256 token
Batch size	16 (addestramento)
Labels	8 (Primary head), neutral, 3 each Secondary head
Tecniche di addestramento	Fine-tuning supervisionato con early stopping (1 epoch per testa primaria, 2 epoch per teste secondarie); bilanciamento delle classi
Metriche di valutazione	Accuratezza, F1-score (weighted), Precisione (weighted), Richiamo (weighted), Matrice di confusione, Curva ROC
Librerie utilizzate	numpy, pandas, sentence-transformers, transformers, torch, scikit-learn, matplotlib, seaborn, data set
Hardware	CPU Intel(R) Core(TM) i7-1165G7, RAM 8 GB, GPU (opzionale, se disponibile)
Framework di addestramento	Hugging Face Transformers Trainer

Meta-classi emozionali

Infine, per valutare al meglio gli errori commessi dal modello e eventuali incongruenze sistematiche nella classificazione, le otto emozioni primarie e le loro variazioni sono state mappate in tre meta-classi (v. sez. 3.3), che si rifanno alle categorie della sentiment analysis [168]: *Positive*, *Neutral* e *Negative*, dove la meta-classe *positive* è stata assegnata alle emozioni

piacevoli, *negative* a quelle sgradevoli e *neutral* a quelle ambigue. I dettagli del mapping sono visibili in tabella 3.5.

Tabella 3.5: Raggruppamento delle emozioni in meta-classi di sentiment

Meta-classe	Emozioni associate
Positive	<i>serenity, joy, ecstasy, acceptance, trust, admiration</i>
Neutral	<i>interest, anticipation, vigilance, distraction, pensiveness, surprise, amazement, annoyance</i>
Negative	<i>sadness, grief, disgust, loathing, boredom, fear, terror, anger, rage, apprehension</i>

La figura 3.8, sintetizza le principali fasi della pipeline di classificazione, dall’input testuale fino all’assegnazione gerarchica dell’emozione primaria e della relativa intensità.

Workflow

L’architettura implementata per la classificazione gerarchica delle emozioni si compone di due livelli distinti, entrambi implementati partendo dal modello sentence-transformers/all-roberta-large-v1, una variante avanzata della famiglia RoBERTa, ottimizzata dalla libreria sentence-transformers [89] per generare rappresentazioni semantiche ad alta densità. Il primo livello, definito testa primaria, classifica un input testuale in una delle otto emozioni primarie definite nel modello di Plutchik nella sezione 2.4. Il secondo livello, invece, sulla base dell’emozione primaria precedentemente predetta, attiva una delle otto teste di classificazione secondarie dedicate. Questo approccio consente di classificare l’emozione secondaria corrispondente, distinguendola in una delle tre possibili intensità associate dal modello di Plutchik all’emozione primaria predetta.

Classificazione delle emozioni primarie

La testa primaria è costituita da un livello lineare che riceve in input l'embedding [CLS] [132] di dimensione 1024 prodotto dal modello di base e lo mappa a 8 classi. L'output del livello lineare è seguito da una funzione di attivazione softmax, implementata tramite `torch.nn.functional.softmax`, che normalizza i valori come probabilità, per fornire una misura di *confidenza* (i.e., probabilità della predizione) relativa del modello. La distribuzione di valori che si ottiene indica quanto è "sicuro" assegnare l'input a ciascuna classe. Il numero totale di parametri di questa testa è calcolato come:

$$1024 \times 8 + 8 = 8200,$$

dove si considerano i pesi della matrice lineare e i termini di bias [146].

Classificazione delle emozioni secondarie

Per la classificazione delle intensità emozionali secondarie, è stato implementato un secondo livello composto da otto teste di classificazione secondarie, una per ciascuna emozione primaria, che classificano le tre classi di intensità. Ciascuna testa secondaria riceve lo stesso embedding [CLS] di dimensione 1024 e lo mappa a tre classi rappresentanti le intensità emozionali specifiche dell'emozione primaria predetta (ad esempio, *ecstasy*, *joy*, *serenity* per *joy*). Analogamente al primo livello, una funzione softmax produce la distribuzione di probabilità che rappresenta la confidenza relativa per ciascuna sottoclasse.

Il numero di parametri per ogni testa secondaria è:

$$1024 \times 3 + 3 = 3075,$$

per un totale di:

$$3075 \times 8 = 24600$$

parametri complessivi per tutte le teste secondarie.

Ogni testa è addestrata separatamente, prima di essere inserita nell’architettura gerarchica del sistema finale, secondo le specifiche definite nel paragrafo relativo al fine-tuning della sezione 2.3.

Pipeline di caricamento, preprocessamento e classificazione

- *Preprocessamento*: il testo in input è tokenizzato [55] e convertito in tensori PyTorch utilizzando il tokenizer di sentence-transformers/all-roberta-large-v1, come implementato nella funzione `preprocess_text`. La tokenizzazione include padding [169] e troncamento a una lunghezza massima di 256 token, producendo i tensori `input_ids` e `attention_mask` (v. sez. 2.3).
- *Caricamento modelli e tokenizer*: il tokenizer e il modello base sentence-transformers/all-roberta-large-v1 sono caricati tramite la funzione `load_models`, utilizzando le API di `transformers.AutoTokenizer` e `transformers.AutoModel`. Il modello è configurato per generare l’embedding `[CLS]` utilizzato dalle teste di classificazione.
- *Caricamento teste di classificazione*: la testa primaria è caricata dal path specifico dove era stata salvata in fase di addestramento, per la classificazione delle 8 emozioni primarie; il tutto è implementato nella funzione `load_heads`. Ciascuna delle otto teste secondarie, una per ogni emozione primaria, è caricata separatamente dal path dove era stata salvata, per le rispettive intensità emozionali.
- *Caricamento encoder di etichette*: un LabelEncoder è utilizzato per tradurre gli indici numerici in etichette testuali per la testa primaria e ciascuna testa secondaria, come implementato nella funzione `load_label_encoders`. Gli encoder sono caricati per garantire coerenza tra le etichette predette e quelle effettive durante l’inferenza.
- *Classificazione*: per ogni frase, si esegue la classificazione primaria uti-

lizzando la funzione `classify_text`, ottenendo le probabilità per ciascuna delle 8 classi, tramite `torch.nn.functional.softmax`. La classe con la confidenza relativa più alta è selezionata come emozione primaria predetta e le prime cinque emozioni sono registrate utilizzando `torch.topk`. La frase, poi, viene passata alla testa secondaria corrispondente per predire l'intensità emozionale.

L'intero processo è implementato utilizzando PyTorch e le librerie *transformers* e *sentence-transformers*. La funzione `torch.nn.functional.softmax` è usata per normalizzare le probabilità di output delle teste di classificazione. Durante l'inferenza [137], `torch.no_grad()` è utilizzato per evitare il calcolo dei gradienti, riducendo il costo computazionale. Il modello opera su un dispositivo *cuda* se disponibile, altrimenti su *cpu*, come specificato nella funzione `classify_text`. In figura 3.5, si propone una visione riassuntiva dell'architettura di classificazione appena descritta.

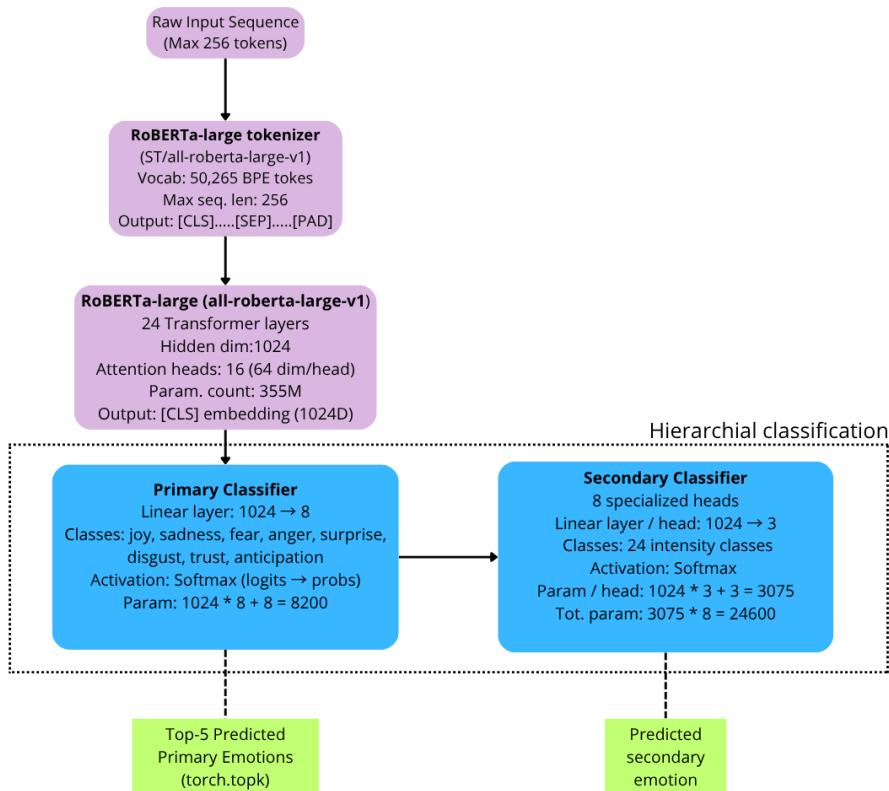


Figura 3.5: Diagramma dettagliato dell'architettura gerarchica

Transfer Learning

Il Transfer Learning (v. sez. 2.3) è stato implementato tramite il framework Hugging Face Transformers Trainer [170], ottimizzando esclusivamente le teste di classificazione mentre i pesi del modello base sono rimasti congelati.

Questo approccio riduce il numero di parametri da ottimizzare (8200 per la testa primaria e 24600 per le teste secondarie), limitando il carico computazionale e preservando le capacità semantiche del modello pre-addestrato. In questo esperimento, sono stati valutati preliminary due distinti approcci di addestramento: nel primo, già descritto, la *primary head* è stata addestrata per la classificazione delle otto emozioni primarie secondo il modello di Plutchik; nel secondo, la *primary head* è stata addestrata per riconoscere le stesse otto emozioni primarie con l'aggiunta di una ulteriore classe *neutral*, al fine di includere anche la possibilità di assenza di emozione specifica.

L'analisi comparativa dei risultati ha evidenziato che l'inclusione della classe *neutral* comporta un decremento di circa 1.5 punti percentuali delle prestazioni di accuratezza nella classificazione primaria rispetto al modello che considera esclusivamente le otto emozioni di base. Tale riduzione della performance a livello primario rischia di propagarsi nelle successive fasi di classificazione secondaria, in quanto l'affidabilità delle *secondary heads* risulta strettamente dipendente dalla correttezza della predizione effettuata dalla *primary head*.

Pertanto, al fine di preservare la massima efficacia e stabilità dell'intero sistema, l'approccio adottato per lo sviluppo definitivo è stato quello che prevede l'addestramento della *primary head* esclusivamente sulle otto emozioni primarie del modello di Plutchik.

1. Fine-tuning della Testa Primaria

La testa primaria è stata addestrata per classificare le nove emozioni primarie; sono stati selezionati 1200 esempi per etichetta. I parametri

di addestramento sono stati configurati per garantire una convergenza stabile e prevenire l’overfitting. Il *learning rate* [156] è stato impostato a 3×10^{-5} , un valore che bilancia la velocità di apprendimento con la stabilità della discesa del gradiente. Il *batch size* [157] è stato definito a 16 per l’addestramento e 32 per la valutazione, ottimizzando l’efficienza computazionale e la stabilità del gradiente. Sono state consentite fino a 12 epoche di addestramento, con *early stopping* [112] attivato dopo un epoch senza miglioramenti nell’accuratezza sul set di validazione, per prevenire l’overfitting e ridurre il tempo di addestramento. Un *weight decay* [158] di 0.001 è stato applicato per penalizzare pesi troppo grandi, migliorando la generalizzazione. L’ottimizzatore utilizzato è AdamW [171], con parametri standard ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), scelto per la sua efficacia in problemi di classificazione profonda. La funzione di perdita è la *cross-entropy loss* [107], standard per la classificazione multiclasse. Il modello migliore, basato sull’accuratezza sul set di validazione, è stato salvato, mantenendo un solo checkpoint per limitare l’occupazione di memoria.

2. Fine-tuning delle Teste Secondarie

Otto teste secondarie sono state addestrate separatamente, ciascuna su un sottoinsieme di dati contenente 1800 esempi (600 per intensità) filtrati per l’emozione primaria corrispondente. I parametri di addestramento sono analoghi a quelli della testa primaria, con due modifiche principali. L’*early stopping* è stato configurato con una pazienza di due epoche, consentendo una maggiore esplorazione dello spazio dei parametri, dato il numero ridotto di classi (tre per testa secondaria rispetto a otto per la testa primaria). Inoltre, è stato implementato il tracciamento della loss ogni 50 step di addestramento, con generazione di grafici della loss per epoch per monitorare la convergenza e identificare eventuali anomalie durante l’addestramento. Come per la

testa primaria, solo i parametri delle teste di classificazione sono stati ottimizzati, mantenendo i pesi del modello base invariati.

Valutazione delle Prestazioni

Le metriche di valutazione, di cui alla sezione 3.3, sono state calcolate separatamente per la testa primaria e per ciascuna testa secondaria. Il processo di valutazione è stato applicato anche al mapping in meta-classi, come descritto nella tabella 3.5 della sezione 3.3, che mappa le emozioni in **positive**, **neutral** e **negative**. Per un'analisi approfondita degli errori, i cinque errori di predizione più frequenti per ciascuna metaclasse sono stati identificati e salvati in un file CSV, consentendo un esame dettagliato delle discrepanze. Le metriche calcolate includono *precision*, *recall* e *F1-score* per valutare la qualità delle predizioni, integrate dall'analisi qualitativa degli errori tramite la funzione *Counter* per quantificare le coppie di etichette errate più comuni. Questo approccio ha permesso una valutazione rigorosa delle prestazioni del modello considerando solo le tre metaclassi appena descritte.

3.4 Classificazione degli Stati Affettivi Complessi

L'obiettivo di questo passo della classificazione secondo il modello di Plutchick è identificare stati affettivi complessi dedotti dalla presenza di coppie di emozioni primarie. Seguendo il modello di Plutchick, dove gli stati affettivi composti da due emozioni si trovano fuori dal fiore, usciamo dalla classificazione gerarchica in favore di un approccio basato sulla similarità semantica tra le emozioni vicine e le frasi da classificare.

Usiamo il modello Sentence-T5 [123] con la metrica FAISS [85], sulla base dei risultati descritti nella sezione 4.1, per valutare la prossimità semantica.

Il processo si articola in due passaggi principali: nel primo, si identificano le coppie di emozioni primarie classificate come top-5, che corrispondono a stati affettivi composti definiti dal modello di Plutchik; nel secondo, si calcola la similarità semantica tramite FAISS tra l'embedding della frase e quello dell'etichetta relativa allo stato affettivo rilevato, accettando la classificazione in tale stato solo se la similarità supera una soglia, che abbiamo definito sperimentalmente. Per top-5 si intende l'insieme delle cinque emozioni primarie con la più alta probabilità predetta dalla testa primaria di classificazione: queste probabilità derivano dalla distribuzione di output dell'ultimo layer del modello (i.e., softmax), che assegna a ciascuna delle emozioni primarie una confidenza. Selezionare le cinque emozioni con il punteggio più elevato consente di limitare l'analisi a quelle con maggiore probabilità di correttezza, riducendo l'impatto del rumore nelle predizioni meno certe. In presenza di multipli stati affettivi composti, l'analisi si restringe alle top-3 emozioni primarie, per filtrare predizioni meno probabili (i.e., con una minore confidenza relativa), valutando la similarità solo per le combinazioni valide.

Le soglie sono state testate sperimentalmente come segue:

1. **Soglia di Similarità (Similarity_Threshold, ST)**: compresa tra 0.60 e 0.75, verifica che la similarità coseno tra l'embedding della frase e quello dell'emozione composta superi il valore specificato;
2. **Soglia di Probabilità (Probability_Threshold, PT)**: compresa tra 0.01 e 0.10, garantisce che la confidenza relativa delle emozioni seconda e terza nella top-3, predette dalla testa primaria, sia sufficientemente elevata;
3. **Soglia di Probabilità Rimanente (Remaining_Prob_Threshold, RPT)**: compresa tra 0.05 e 0.15, filtra le seconda e terza

emozione nella top-3, in base alla probabilità composta, ossia dalla confidenza relativa residua calcolata come:

$$P_{\text{comp}}(e_1, e_2) = \frac{P(e_2)}{1 - P(e_1)} \quad (3.1)$$

dove:

- e_1 : emozione primaria predetta con la probabilità più alta;
- e_2 : seconda emozione tra le top-5, candidata a formare una combinazione affettiva con e_1 ;
- $P(e_1)$: probabilità assegnata all'emozione primaria e_1 ;
- $P(e_2)$: probabilità assegnata all'emozione e_2 ;
- $P_{\text{comp}}(e_1, e_2)$: probabilità normalizzata della combinazione affettiva composta da e_1 e e_2 , calcolata come rapporto tra la probabilità di e_2 e la probabilità residua (non coperta da e_1).

Lo stesso calcolo viene ripetuto per entrambe le predizioni diverse dalla principale nella top-3. L'esperimento si pone come scopo quello di individuare la combinazione e il valore ottimale delle soglie citate, così da poterlo applicare al modello finale garantendo la maggiore accuratezza possibile nel rilevamento degli stati affettivi.

Durante questo esperimento, sono state valutate varie combinazioni di parametri, andando ad escludere le predizioni poco rilevanti. I parametri testati sono elencati in tabella 3.6.

Tabella 3.6: Parametri sperimentali per la classificazione degli stati affettivi composti

Parametro	Valore
Modello linguistico	sentence-transformers/all-roberta-large-v1 (V.paragrafo 3.3), sentence-transformers/sentence-t5-base (V. paragrafo3.1)
data set	final_emotion_data set (parte 2)
Lunghezza massima sequenze	256 token
Batch size	Non specificato
Labels	8 (stati affettivi composti: Love, Submission, Awe, Disapproval, Remorse, Contempt, Aggressiveness, Optimism)
Tecniche di valutazione	Classificazione degli stati affettivi composti basata su coppie di emozioni primarie (top-5 e top-3) con similarità semantica FAISS; teste pre-addestrate caricate da Models_9_finetuning/all-roberta-large-v1-finetuned e models_{emo}/Sec_{emo}-finetuned
Metriche di valutazione	Accuratezza, Precisione (macro), Richiamo (macro), F1-score (macro), Matrice di confusione
Librerie utilizzate	<code>torch</code> , <code>torch.nn.functional</code> , <code>numpy</code> , <code>pandas</code> , <code>scikit-learn</code> , <code>transformers</code> , <code>sentence-transformers</code> , <code>matplotlib</code> , <code>faiss</code>
Hardware	CPU Intel(R) Core(TM) i7-1165G7, RAM 8 GB, GPU (opzionale, se disponibile via CUDA)
Framework di valutazione	Hugging Face Transformers
Soglie di valutazione	Soglie di similarità: [0.60, 0.65, 0.70, 0.75], Soglie di probabilità: [0.01, 0.05, 0.10]
Metrica di similarità	FAISS (IndexFlatIP) per similarità semantica

Workflow

In questo esperimento abbiamo due flussi di lavoro per valutare due approcci distinti, basati su un’architettura che integra classificazione delle emozioni primarie e rilevazione di stati affettivi composti tramite similarità semantica. L’architettura è simile a quella per le emozioni primarie e secondarie descritta in precedenza (v. sezione 3.3); con l’aggiunta della logica di rilevamento degli stati affettivi composti: le coppie di emozioni top-5 sono confrontate con un dizionario predefinito di stati affettivi composti, e.g., Love= (“joy”, “trust”). La similarità tra l’embedding della frase e quelli degli stati affettivi composti candidati è calcolata selezionando i candidati che superano le **soglie di similarità** testate, ossia $similarity_threshold = 0.60, 0.65, 0.70, 0.75$.

Un’analisi statistica preliminare condotta sul data set (v. sez. 2.5) mostra che la similarità calcolata non scende mai sotto il valore minimo osservato di 0.6848. Tale osservazione consente di ipotizzare che valori inferiori siano indicativi di una bassa o assente correlazione semantica. Pertanto, nella definizione delle soglie operative per la fase di classificazione affettiva composta, sono state selezionate soglie comprese tra 0.60 e 0.75, per investigarne l’impatto sull’accuratezza del modello, mantenendo comunque la soglia minima empirica osservata come riferimento per la discriminabilità tra stati affettivi rilevati e non. Nel **primo flusso**, la validazione degli stati affettivi composti si basa esclusivamente su questa soglia. Per evitare falsi positivi in contesti di frase ambigui, è stato dunque impostato il **secondo flusso**, con un vincolo aggiuntivo: le emozioni primarie che formano la coppia candidata devono superare una soglia di probabilità, i.e. confidenza relativa, ossia $prob_threshold = 0.01, 0.05, 0.10$ e una soglia di probabilità condizionata $remaining_prob_threshold = 0.05, 0.10, 0.15$, calcolata come da formula [3.1], definita nella sezione 3.4. Entrambe le soglie di probabilità sono state definite attraverso analisi preliminari condotte sulla porzione di data set non impiegata per la valutazione (parte 1 2.5). In particolare, per gli esempi caratterizzati da uno stato affettivo composto, è stato calcolato il valore minimo della probabilità associata alla predizione principale, definito come $prob_threshold$ e pari a 0.03, nonché la soglia per la probabilità dinamica delle altre due emozioni incluse tra le $top-3$, indicata come $remaining_prob_threshold$ e pari a 0.07. Tali valori soglia sono stati successivamente raffinati esplorando intervalli prossimi ai minimi rilevati, al fine di ottimizzare la sensibilità del modello nei casi più ambigui e stratificati. Questo vincolo basato sulla probabilità condizionata rispetto all’emozione primaria predetta permette di normalizzare i valori su cui confrontare le altre due emozioni in $top-3$. In caso di assenza di stati affettivi composti nelle $top-5$, si restituisce “none”; con un solo stato affettivo composto, questo sarà validato se supera la soglia di similarità; con multipli stati affettivi

composti, si analizzano le top-3 per verificare quali soddisfino le soglie di confidenza relativa e condizionata.

Il primo approccio elimina l'applicazione della soglia di probabilità, valutando esclusivamente la similarità semantica tra lo stato affettivo composto rilevato e l'etichetta teorica di Plutchik corrispondente; la predizione viene accettata solo qualora il valore di similarità superi la soglia prestabilita (*similarity threshold*).

Il secondo approccio, invece, valuta l'impatto congiunto delle due soglie descritte: da un lato, il livello di confidenza associato alla predizione della testa primaria relativamente alle emozioni di base che costituiscono lo stato affettivo (*probability threshold*); dall'altro, il valore di similarità semantica rispetto all'etichetta target dello stato affettivo composto (*similarity threshold*).

L'approccio che, a seguito di questa valutazione, fornirà le prestazioni di accuratezza più elevate nel rilevamento degli stati affettivi, misurate confrontando le etichette predette con quelle effettivamente presenti nel data set descritto nella Sezione 2.5, sarà selezionato come configurazione finale del modello.

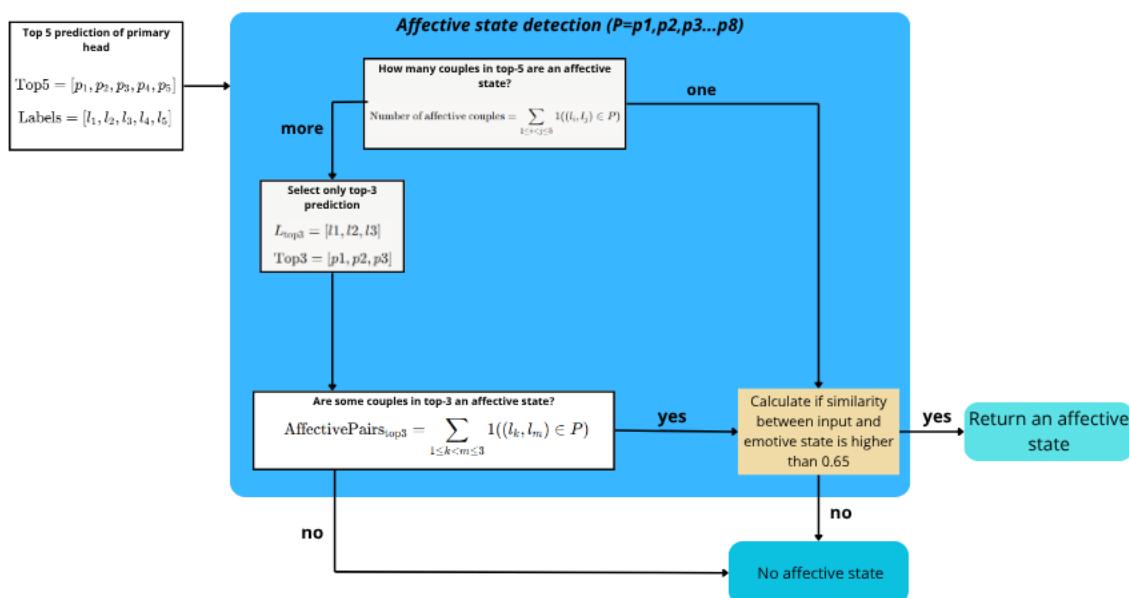


Figura 3.6: Approccio (1): Flusso con solo controllo sulla soglia di similarità.

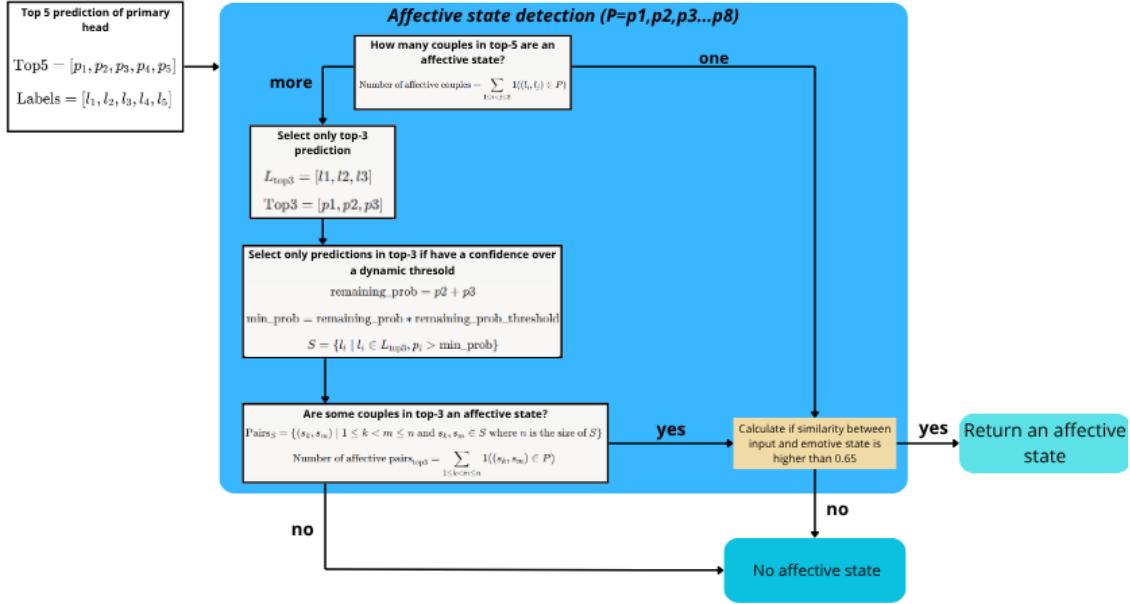


Figura 3.7: Approccio (2): Flusso con controllo aggiuntivo sulla probabilità delle emozioni primarie predette in Top-3.

3.5 Applicazione al Data set di Grooming

Il nostro sistema di emotion detection è stato infine applicato ad ogni *sentence* del data set di grooming. Per ogni istanza (i.e., riga), il sistema ha generato i tre livelli di classificazione emotiva:

- *primary_emotion*: emozione primaria tra le 8 classi emozionali o label *neutral*
- *secondary_emotion*: tre sottocategorie relative al livello di arousal per ogni emozione primaria.
- *composed_emotion*: eventuali stati affettivi composti associati a due emozioni vicine.

Per analizzare i pattern emotivi nelle varie fasi del grooming (v. sez. 1.1), abbiamo valutato sette metriche di similarità, ciascuna con specifica utilità analitica:

1. **Frequenze Relative** ($p_i = \frac{\text{Conteggio emozione}_i}{\text{Totale osservazioni}}$) [172]: utili per identifica-

re le emozioni predominanti in ciascuna fase di grooming, permettono di quantificare la distribuzione delle risposte emotive e confrontare direttamente l'incidenza tra diversi tag/fasi.

Range: $p_i \in [0, 1]$ (con $\sum p_i = 1$ su tutte le emozioni).

2. **Indice di Dominanza** ($D = \frac{\text{Freq. emozione dominante}}{\text{Totale}}$) [173]: misura la concentrazione emotiva in termini di frequenza, utile per individuare strategie di manipolazione emotiva focalizzata su specifiche emozioni (valori alti) o approcci emotivamente diversificati (valori bassi).

Range: $D \in [\frac{1}{N}, 1]$, dove N è il numero di emozioni; più alto D , maggiore la dominanza.

3. **Entropia di Shannon** ($H = -\sum p_i \log_2 p_i$) [174]: quantifica la complessità dei pattern emotivi, indipendentemente dall'emozione stimolata. Bassi valori (1.5-2 bit) indicano strategie emotive coerenti, che potremmo associare ipoteticamente alle fasi iniziali di costruzione della fiducia, mentre valori alti (2.5-3 bit) potrebbero essere associati a tattiche complesse, volutamente destabilizzanti.

Range: $H \in [0, \log_2 N]$, con N numero di emozioni.

4. **Probabilità di Transizione** ($P(E_{t+1} | E_t)$) [175]: modella la dinamica temporale delle emozioni, rivelando sequenze potenzialmente predittive. Utile per l'identificazione di pattern sequenziali caratteristici nel grooming. Si basa sulla probabilità condizionata [176] tra emozioni consecutive, calcolata approssimando i conteggi osservati nel data set:

$$P(E_{t+1} | E_t) = \frac{\#(E_t, E_{t+1})}{\#(E_t)}$$

dove $\#(E_t, E_{t+1})$ è il numero di occorrenze della coppia di emozioni tra due fasi o tag consecutivi e $\#(E_t)$ è il numero di occorrenze dell'emozione E_t nella fase o tag corrente.

Range: $P \in [0, 1]$ (ogni transizione ha una probabilità condizionata).

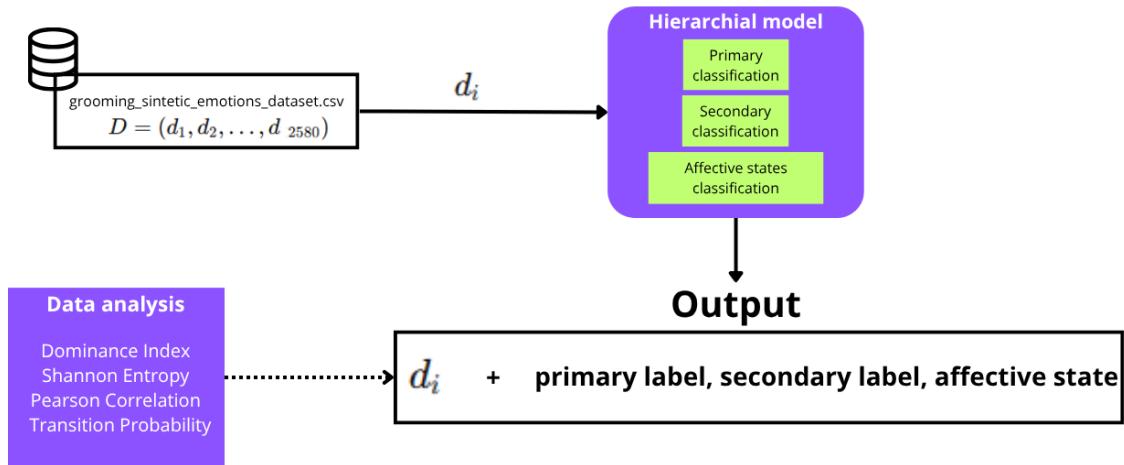


Figura 3.8: Workflow della classificazione finale

Capitolo 4

Risultati e conclusioni

In questo capitolo riassumiamo in modo organico i principali risultati emersi nel corso di questa tesi di ricerca, analizzandone l'impatto scientifico e le ricadute pratiche lungo l'intero processo sperimentale.

L'intero processo ha seguito un approccio metodologico in cui ogni fase sperimentale è stata progettata per rispondere a un quesito specifico, i cui risultati hanno guidato in maniera diretta la configurazione e l'impostazione dell'esperimento successivo. Tale strategia ha permesso una progressiva costruzione e raffinamento del sistema, assicurando coerenza tra le scelte tecniche e le evidenze raccolte.

Discutendo criticamente i dati ottenuti, metteremo in luce i punti di forza dell'approccio adottato, le criticità riscontrate e le potenziali direzioni per sviluppi futuri.

4.1 Risultati Embedding e Prossimità Semantica

Nella Tabella 4.1 sono riportati i valori di accuratezza ottenuti per k compreso tra 1 e 3. I risultati relativi a $k = 4$ e $k = 5$, seppur calcolati durante lo studio, sono stati esclusi dalla rappresentazione in quanto confermano

pienamente l’andamento già osservato, senza apportare variazioni sostanziali alle conclusioni.

Il modello *Sentence-T5*, ad esempio, si distingue per ottenere i valori di accuratezza più elevati per ciascuna metrica considerata nei casi $k = 1, 2, 3$. Tale tendenza si mantiene anche per $k = 4$ e $k = 5$: per $k = 4$, si osservano valori pari a 0,8244 per la *cosine similarity base* [75] (contro un massimo di 0,7893 tra gli altri modelli), 0,8484 con *FAISS* [85] (contro 0,8216), e 0,8318 con la metrica di *Roger-Tanimoto* [82] (contro 0,8178). Analogamente, per $k = 5$, le rispettive accuratezze sono 0,8344, 0,8480 e 0,8331, sempre superiori rispetto a quelle ottenute dagli altri modelli (0,7924, 0,8242 e 0,8178 rispettivamente).

Questi risultati confermano l’andamento delle prestazioni del modello *Sentence-T5* [123], già evidente per $k = 1, 2, 3$ e giustificano la scelta di limitare la trattazione ai primi tre valori di k , sufficienti a evidenziare con chiarezza il comportamento dei modelli analizzati rispetto al compito in esame.

Tabella 4.1: Accuracy [121] dei modelli su *emotion_label2* (9 emozioni) con Cosine Similarity (CS), FAISS Cosine Similarity (FS) e Roger Tanimoto (RT).

Modello	k=1			k=2			k=3		
	CS1	FS1	RT1	CS2	FS2	RT2	CS3	FS3	RT3
paraphrase-mpnet	0.8056	0.8056	0.7987	0.7267	0.8056	0.7987	0.7751	0.8053	0.7940
multilingual-paraphrase-mpnet	0.7836	0.7836	0.7736	0.6993	0.7836	0.7736	0.7567	0.7827	0.7687
roberta	0.8031	0.8031	0.8000	0.7329	0.8031	0.8000	0.7918	0.8104	0.7998
distilbert	0.7722	0.7722	0.7658	0.6973	0.7722	0.7658	0.7589	0.7807	0.7756
bert-base-uncased	0.6364	0.6364	0.5462	0.5396	0.6364	0.5462	0.5829	0.6358	0.5640
all-MiniLM-L6-v2	0.7273	0.7273	0.6907	0.6500	0.7273	0.6907	0.7004	0.7378	0.7091
all-MPNet-base-v2	0.7993	0.7993	0.7869	0.7156	0.7993	0.7869	0.7738	0.8027	0.7949
paraphrase-distilroberta-base-v1	0.7924	0.7924	0.7953	0.7164	0.7924	0.7953	0.7722	0.8002	0.8011
multi-qa-mpnet-base-dot-v1	0.8064	0.8064	0.7949	0.7413	0.8064	0.7949	0.7947	0.8160	0.8109
stsbert-large	0.7602	0.7602	0.7529	0.6851	0.7602	0.7529	0.7442	0.7684	0.7580
distiluse-base-multilingual-cased	0.7547	0.7547	0.7293	0.6558	0.7547	0.7293	0.7247	0.7567	0.7407
sentence-t5-large	0.8369	0.8369	0.8131	0.7760	0.8369	0.8131	0.8229	0.8433	0.8238

In Tabella 4.1 e nelle Figure 4.1, 4.2 e 4.3 di seguito, per il task specifico

di *semantic similarity* applicato alla classificazione emozionale, notiamo che la metrica *FAISS Cosine Similarity*(FS) risulta preferibile rispetto alla *Cosine Similarity classica* (CS) e alla *Roger Tanimoto* (RT). La metrica CS mostra un degrado significativo delle prestazioni con l'aumento di k , come evidenziato ad esempio da *sentence-t5-large* che passa da 0.8369 a $k = 1$ a 0.7760 a $k = 2$ e da *multi-qa-mpnet-base-dot-v1* [143] che scende da 0.8064 a 0.7413 nello stesso intervallo.

Al contrario, FS mantiene accuracy elevate e stabili, raggiungendo valori come 0.8433 a $k = 3$ per *sentence-t5-large* e 0.8160 per *multi-qa-mpnet-base-dot-v1*, con prestazioni diffusamente superiori per $k \geq 1$ rispetto a CS.

La metrica RT, pur mostrando un miglioramento con l'aumento di k (ad esempio, 0.8238 a $k = 3$ per *sentence-t5-large* e 0.8109 per *multi-qa-mpnet-base-dot-v1*), non supera FS nei casi migliori e richiede valori di k più elevati per raggiungere il picco massimo, a testimonianza di un'analisi della similarità non sempre precisa.

La preferenza per FS è quindi motivata dalla sua robustezza numerica, dalla capacità di gestire efficacemente la similarità semantica senza introdurre rumore con l'aumento del numero di vicini, mantenendo una accuracy simile per tutti i valori di k e infine dalle prestazioni ottimali su più modelli come *sentence-t5-large* e *multi-qa-mpnet-base-dot-v1*, rendendola diffusamente la scelta migliore per questo compito.

Dall'analisi quantitativa riportata nei grafici 4.1, 4.2 e 4.3 il modello *sentence-t5-large* si conferma come il migliore in termini di accuratezza su tutte e tre le metriche di similarità semantica considerate: *Cosine Similarity* (CS), *FAISS Cosine Similarity*(FS) e *Roger Tanimoto* (RT) e per ciascuno dei valori di k (1, 2 e 3). Le sue performance sono costantemente superiori rispetto a tutti gli altri modelli confrontati, raggiungendo valori di accuratezza superiori all'83% con CS e oltre l'84% con FS al valore $k=3$. Questa superiorità indica

una migliore capacità di rappresentare semanticamente le frasi associate alle emozioni. Il distacco netto, spesso superiore a 3–5 punti percentuali rispetto ai modelli concorrenti, suggerisce che *sentence-t5-large* cattura in modo più efficace le sfumature semantiche complesse e pertanto risulta la scelta migliore per il nostro task.

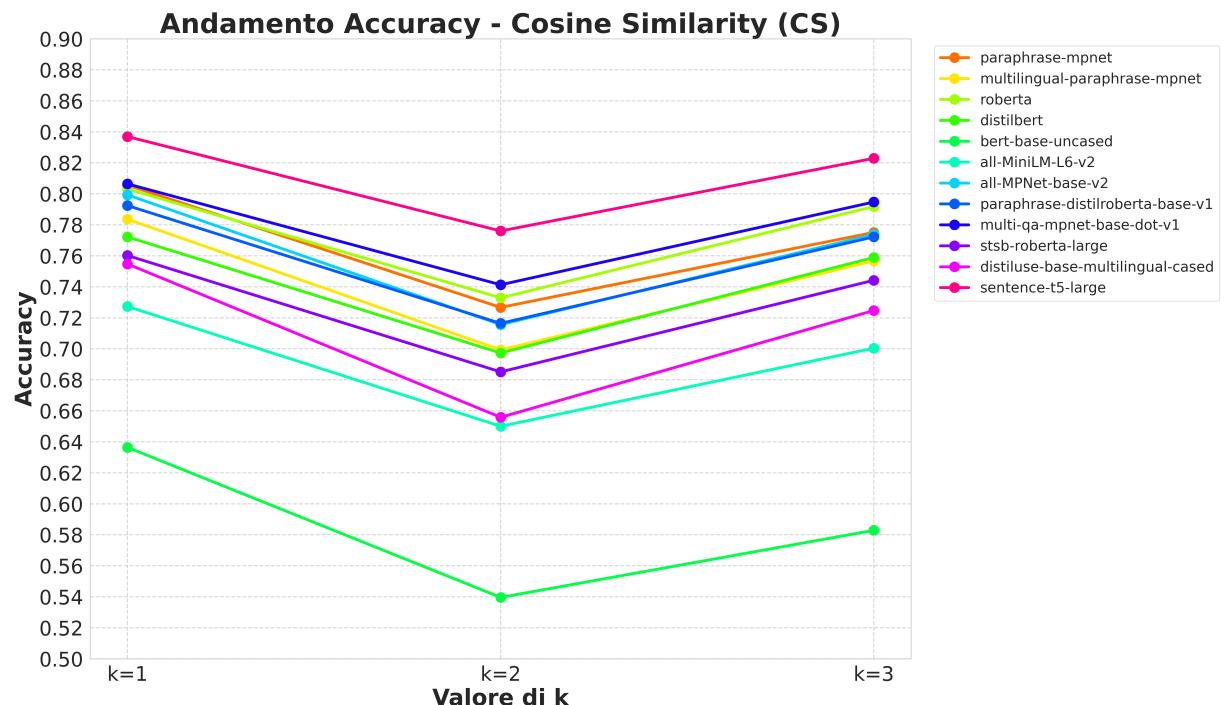


Figura 4.1: Accuracy semantic similarity con CS

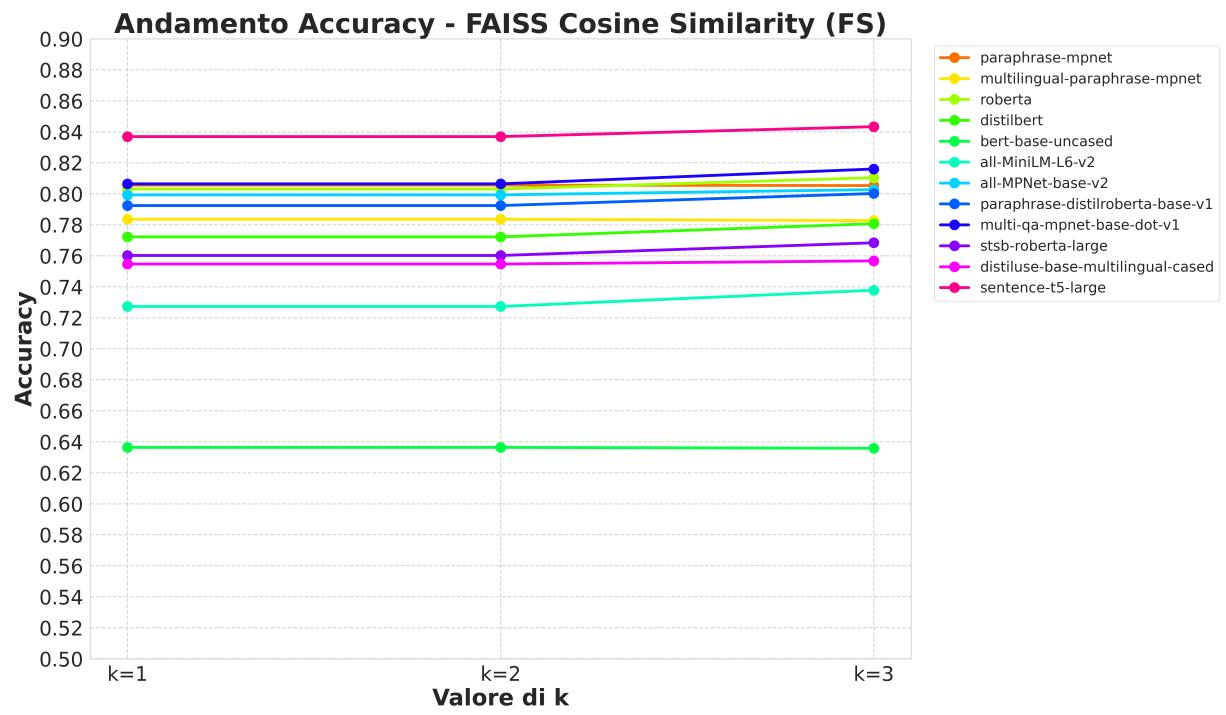


Figura 4.2: Accuracy semantic similarity con FS

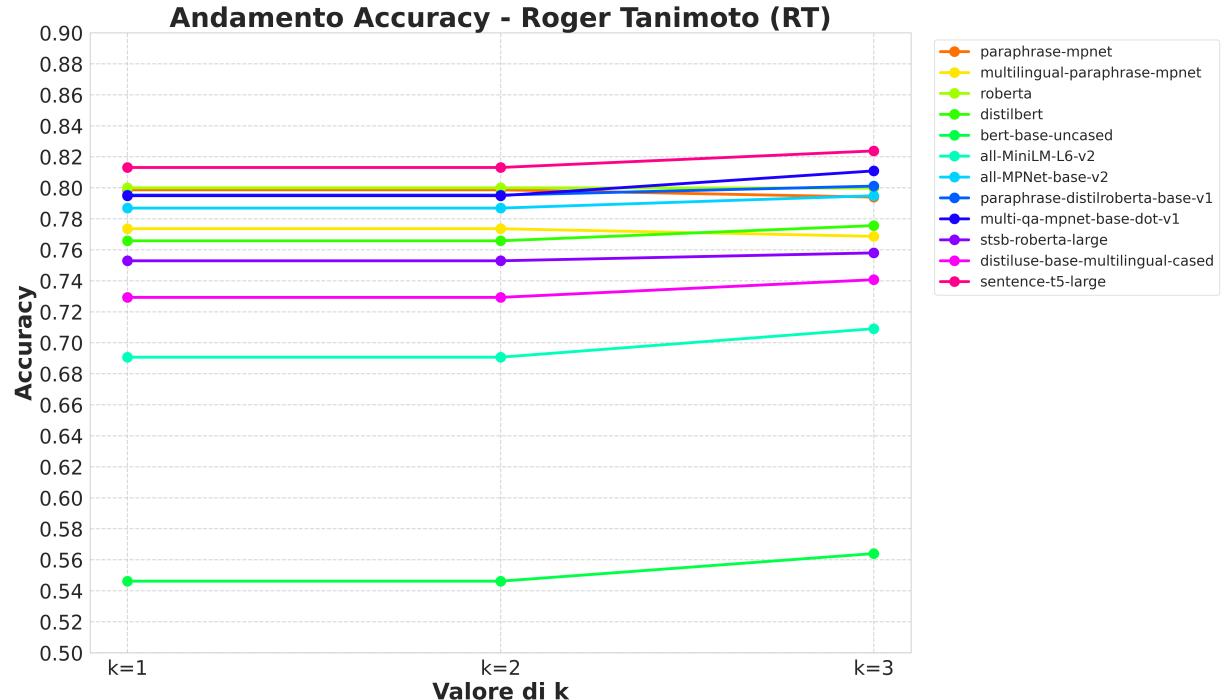


Figura 4.3: Accuracy semantic similarity con RT

L’analisi delle matrici di confusione relative al modello *sentence-t5-large*, rilevato come il più performante tra quelli esaminati, evidenziano un comportamento coerente con l’effetto dell’incremento del parametro k nel task di classificazione emozionale basato su *semantic similarity* [44], come riportato nella Tabella 4.1.

Le Figure 4.4 e 4.5 illustrano le matrici di confusione per FS, evidenziando stabilità nella classificazione, mentre le Figure 4.6 e 4.7 riportano gli errori con CS, mostrando maggiore confusione tra emozioni simili (i.e., vicine nel modello di Plutchick [40]) oppure appartenenti alla stessa metaclasse (e.g., *trust* e *joy*, della metaclasse Positive, v sez. 3.3) a $k \geq 2$. Nella prima coppia di matrici di confusione (Figura 4.4 e Figura 4.5) si osserva un generale equilibrio nelle distribuzioni, con differenze marginali nei valori sulla diagonale principale: in alcune classi la matrice per $k=1$ mostra un numero lievemente superiore di predizioni corrette, mentre in altre classi accade il contrario. Questo indica che, in questa configurazione, l’aumento del parametro k da 1 a 3 non produce variazioni sistematiche significative nella qualità predittiva complessiva.

Al contrario, nella seconda coppia (Figura 4.6 e Figura 4.7), si osserva un chiaro peggioramento delle performance all’aumentare di k . In particolare, la matrice corrispondente a $k=3$ (Figura 4.6) presenta valori sulla diagonale principale sistematicamente inferiori rispetto alla matrice di $k=1$, evidenziando una riduzione della capacità discriminativa del modello con l’inclusione di un numero maggiore di vicini. Questo comportamento suggerisce che l’incremento di k tende nel caso della CS, a introdurre rumore nella decisione finale, cosa che non accade con FS. Per RT, matrici analoghe confermano il miglioramento a $k = 3$, ma FS rimane preferibile per la sua accuracy massima e stabilità. Questi risultati sottolineano la superiorità di FS per il task, grazie alla sua capacità di gestire la *similarità semantica*, riducendo gli errori di classificazione rispetto a CS e offrendo prestazioni leggermente migliori rispetto a RT.

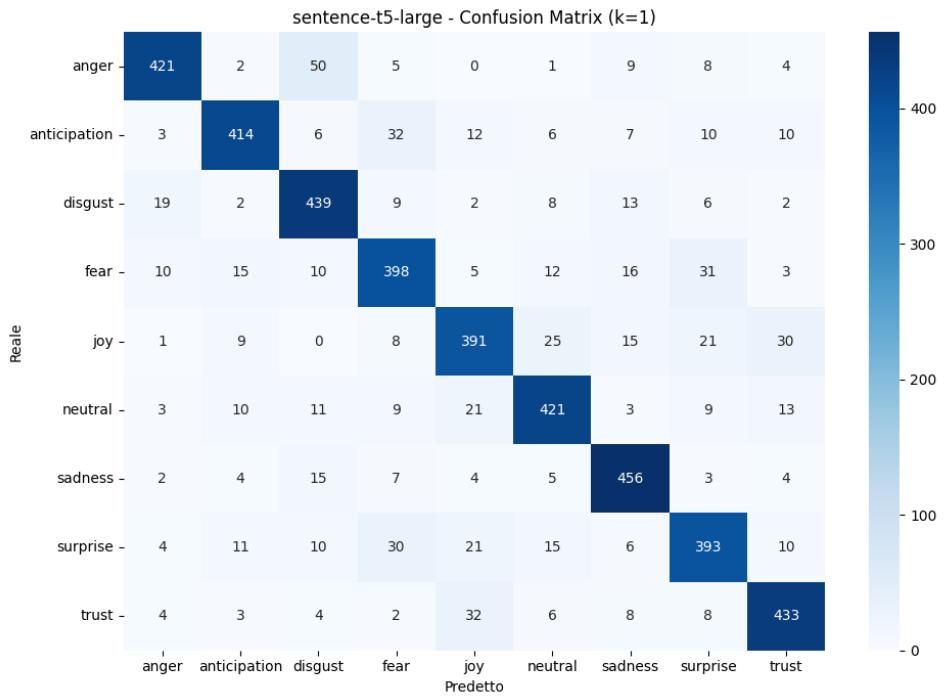


Figura 4.4: Matrice di confusione di Sentence-T5 con $k = 1$ e FS

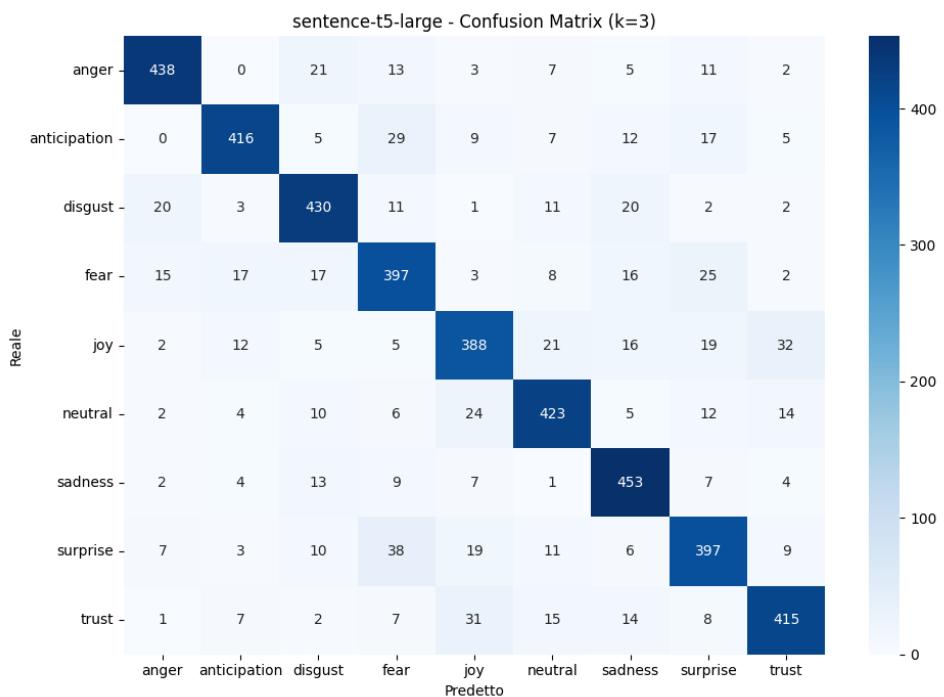


Figura 4.5: Matrice di confusione di Sentence-T5 con $k = 3$ e FS

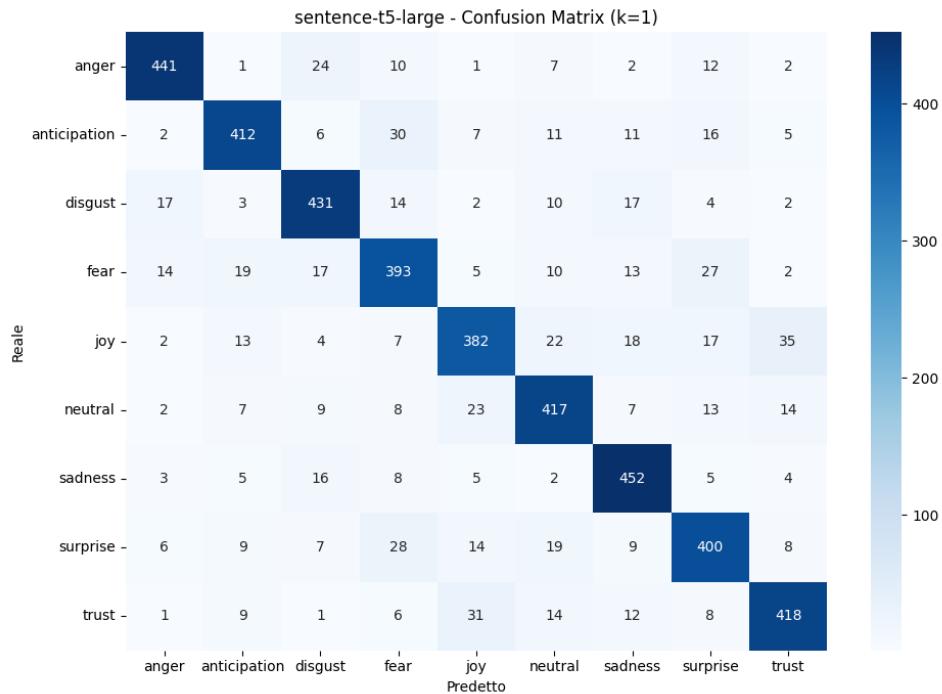


Figura 4.6: Matrice di confusione di Sentence-T5 con $k = 1$ e CS

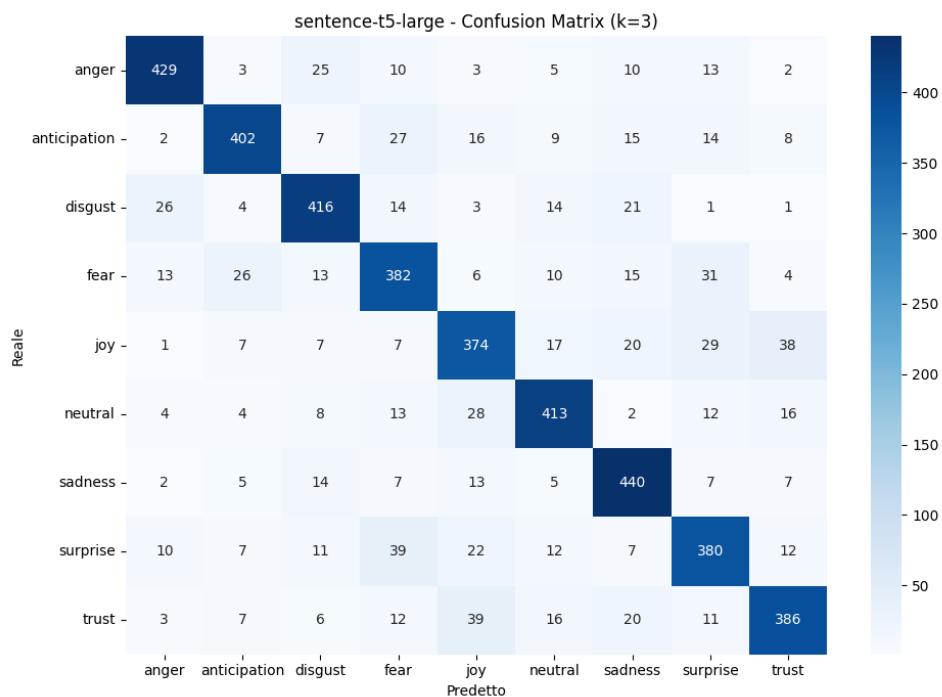


Figura 4.7: Matrice di confusione di Sentence-T5 con $k = 3$ e CS

In virtù dei risultati preliminari del primo test, si è scelto di utilizzare il modello **sentence-T5** e la metrica **FAISS** per il calcolo della similarità tra

embeddings ed emozioni nel modello finale.

4.2 Risultati del Fine-tuning per le Emozioni Primarie

I risultati ottenuti nell’Esperimento 4 si concentrano sull’efficacia dei modelli fine-tuned per la classificazione, esplorando in particolare le prestazioni su un insieme di emozioni, seguendo il modello di Plutchik. Il focus si è spostato sulle differenze prestazionali tra l’approccio di fine-tuning e i metodi precedentemente utilizzati e analizzati nella sezione 4.1, cercando di identificare i modelli in grado di classificare nel migliore dei modi, l’etichetta emozionale associata a ciascun input fornito.

Model	LR	BS	WD	Acc
all-MPNet-base-v2	2E-5	16	0.001	0.65
all-MPNet-base-v2	2E-5	16	0.01	0.65
all-MPNet-base-v2	2E-5	32	0.001	0.63
all-MPNet-base-v2	2E-5	32	0.01	0.65
all-MPNet-base-v2	3E-5	16	0.001	0.67
all-MPNet-base-v2	3E-5	16	0.01	0.68
all-MPNet-base-v2	3E-5	32	0.001	0.65
all-MPNet-base-v2	3E-5	32	0.01	0.65
all-MiniLM-L6-v2	2E-5	16	0.001	0.20
all-MiniLM-L6-v2	2E-5	16	0.01	0.19
all-MiniLM-L6-v2	2E-5	32	0.001	0.19
all-MiniLM-L6-v2	2E-5	32	0.01	0.20
all-MiniLM-L6-v2	3E-5	16	0.001	0.25
all-MiniLM-L6-v2	3E-5	16	0.01	0.24
all-MiniLM-L6-v2	3E-5	32	0.001	0.24

Model	LR	BS	WD	Acc
all-MiniLM-L6-v2	3E-5	32	0.01	0.28
bert-base-uncased	2E-5	16	0.001	0.27
bert-base-uncased	2E-5	16	0.01	0.27
bert-base-uncased	2E-5	32	0.001	0.25
bert-base-uncased	2E-5	32	0.01	0.22
bert-base-uncased	3E-5	16	0.001	0.22
bert-base-uncased	3E-5	16	0.01	0.32
bert-base-uncased	3E-5	32	0.001	0.28
bert-base-uncased	3E-5	32	0.01	0.27
distilbert	2E-5	16	0.001	0.67
distilbert	2E-5	16	0.01	0.67
distilbert	2E-5	32	0.001	0.63
distilbert	2E-5	32	0.01	0.65
distilbert	3E-5	16	0.001	0.70
distilbert	3E-5	16	0.01	0.71
distilbert	3E-5	32	0.001	0.68
distilbert	3E-5	32	0.01	0.67
multi-qa-mpnet-base	2E-5	16	0.001	0.72
multi-qa-mpnet-base	2E-5	16	0.01	0.73
multi-qa-mpnet-base	3E-5	32	0.01	0.72
multi-qa-mpnet-base	2E-5	16	0.01	0.67
multi-qa-mpnet-base	2E-5	32	0.01	0.71
multi-qa-mpnet-base	3E-5	16	0.001	0.75
multi-qa-mpnet-base	3E-5	16	0.01	0.75
multi-qa-mpnet-base	3E-5	32	0.001	0.72
ml-paraphrase-mpnet	2E-5	16	0.001	0.64
ml-paraphrase-mpnet	2E-5	16	0.01	0.64
ml-paraphrase-mpnet	2E-5	32	0.001	0.67
ml-paraphrase-mpnet	2E-5	32	0.01	0.63

Model	LR	BS	WD	Acc
ml-paraphrase-mpnet	3E-5	16	0.001	0.67
ml-paraphrase-mpnet	3E-5	16	0.01	0.67
ml-paraphrase-mpnet	3E-5	32	0.001	0.65
ml-paraphrase-mpnet	3E-5	32	0.01	0.65
paraphrase-mpnet	2E-5	16	0.001	0.65
paraphrase-mpnet	2E-5	16	0.01	0.64
paraphrase-mpnet	2E-5	32	0.001	0.62
paraphrase-mpnet	2E-5	32	0.01	0.63
paraphrase-mpnet	3E-5	16	0.001	0.67
paraphrase-mpnet	3E-5	16	0.01	0.67
paraphrase-mpnet	3E-5	32	0.001	0.65
paraphrase-mpnet	3E-5	32	0.01	0.66
roberta	2E-5	16	0.001	0.77
roberta	2E-5	16	0.01	0.77
roberta	2E-5	32	0.01	0.75
roberta	3E-5	16	0.001	0.80
roberta	3E-5	16	0.01	0.79
stsberta-large	2E-5	16	0.001	0.67
stsberta-large	2E-5	16	0.01	0.67
stsberta-large	2E-5	32	0.001	0.65
stsberta-large	2E-5	32	0.01	0.64
stsberta-large	3E-5	16	0.001	0.70
stsberta-large	3E-5	16	0.0	0.70
stsberta-large	3E-5	32	0.001	0.67
stsberta-large	3E-5	32	0.01	0.68

Legenda:

- **LR:** Learning Rate [156]
- **BS:** Batch Size [157]

- **WD**: Weight Decay [158]
- **Acc**: Evaluation Accuracy [121]

I risultati, riportati in Tabella 4.2, mostrano i valori di accuracy ottenuti per diversi modelli linguistici preaddestrati sottoposti a fine-tuning, variando gli iperparametri chiave come *learning rate*, *batch size* e *weight decay*.

In generale, i modelli che hanno raggiunto le prestazioni migliori sono stati *all-roberta-large-v1* [122] e *multi-qa-mpnet-base-dot-v1*, con un’accuracy massima pari rispettivamente a 0.80 e 0.75. In particolare, *RoBERTa* ha raggiunto l’accuracy più alta con un *learning rate* di $3 \cdot 10^{-5}$, *batch size* di 16 e *weight decay* di 0.001.

Anche *distilbert* [60] ha avuto buone prestazioni, con un’accuracy fino a 0.71 con un *learning rate* di $3 \cdot 10^{-5}$ e *batch size* 16, mostrando che modelli più leggeri possono comunque fornire buoni risultati se ben ottimizzati.

Al contrario, *all-MiniLM-L6-v2* [140] ha mostrato performance inferiori, con accuracy comprese tra 0.19 e 0.28, suggerendo una scarsa capacità del modello nel catturare le caratteristiche rilevanti del task in esame.

I modelli basati su *MPNet* [143], come *all-MPNet-base-v2* e *paraphrase-mpnet*, hanno mostrato risultati mediamente buoni, con accuracy tra 0.63 e 0.68. La variante *multi-qa-mpnet-base-dot-v1*, progettata per task di tipo retrieval [133], ha ottenuto risultati sensibilmente migliori, probabilmente grazie a uno spazio di embedding più strutturato.

I modelli *bert-base-uncased* [60] e *stsbert-roberta-large* [122] hanno mostrato un comportamento più variabile: per *bert-base-uncased*, le accuracy si sono mantenute tra 0.22 e 0.32, mentre per *stsbert-roberta-large* si è osservato un picco a 0.70, ma con maggior variabilità rispetto a *roberta* o *distilbert*.

I modelli *multilingual-paraphrase-mpnet* hanno mostrato risultati di poco inferiori, con una massima di accuracy di 0.67, ma con una maggiore sensibilità alla variazione di *batch size* e *learning rate*.

Complessivamente, i risultati evidenziano come la scelta del modello di partenza e l'ottimizzazione degli iperparametri influiscano in modo significativo sulla performance finale.

L'accuracy tende a migliorare con *learning rate* pari a $3 \cdot 10^{-5}$ e *batch size* contenuti (16), mentre il *weight decay* ha un impatto meno marcato ma comunque rilevante nei modelli più performanti.

Di seguito, vengono riportate le matrici di confusione relative alle due combinazioni che offrono la massima accuracy, entrambe relative al modello *RoBERTa*, con settaggi specifici rispettivamente di LR a 3.00E-05, BS a 16 e WD a 0.001 nel primo caso [Fig. 4.8] e LR a 3.00E-05, BS a 16 e WD a 0.01 nel secondo [Fig. 4.9]. Come si nota, gli errori nella classificazione sono molto limitati e circoscritti a label che rappresentano emozioni vicine semanticamente (i.e., petali vicini nel modello di Plutchick, e.g., disgust-anger, anticipation-joy).

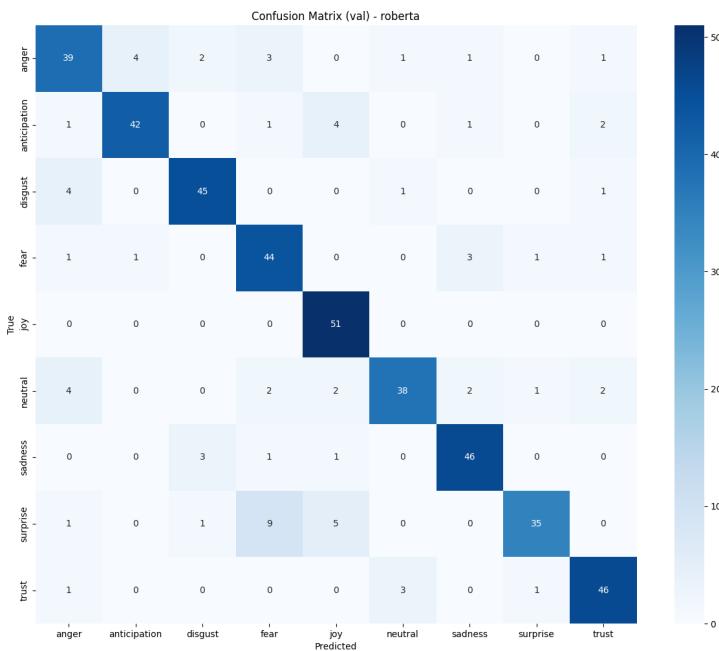


Figura 4.8: Matrice di confusione di RoBERTa con LR a 3.00E-05, BS a 16 e WD a 0.001

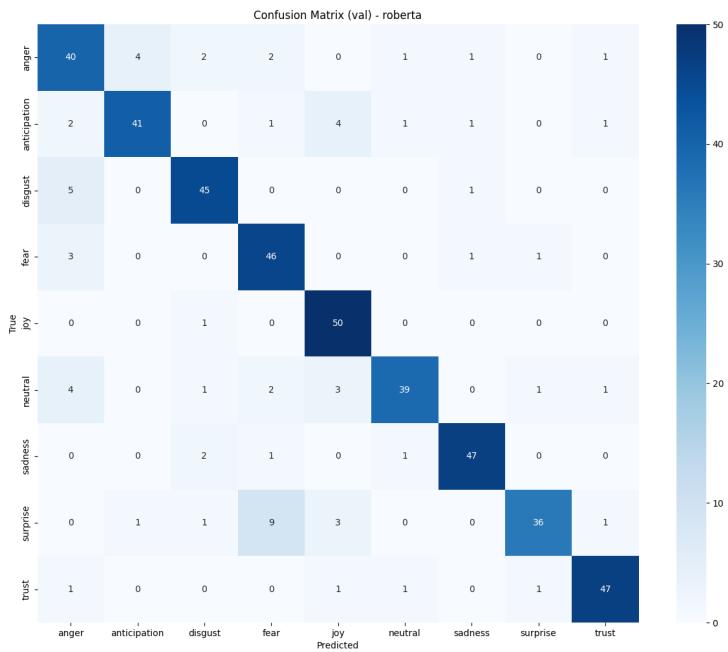


Figura 4.9: Matrice di confusione di RoBERTa con LR a 3.00E-05, BS a 16 e WD a 0.01

Avendo identificato i modelli e gli iperparametri più performanti, per gli esperimenti successivi ci siamo concentrati sul fine-tuning del modello *RoBERTa*, risultato il migliore con un netto divario di accuracy rispetto a tutti gli altri.

4.3 Risultati Addestramento delle Teste di Classificazione

RoBERTa si è dimostrato l’architettura più performante per il task di fine-tuning sui dati in esame. Sono state dunque sviluppate e integrate teste di classificazione specializzate, ciascuna ottimizzata sulla base della combinazione di iperparametri che ha mostrato le migliori prestazioni negli esperimenti, che costituiscono il nucleo del sistema di classificazione gerarchica finale.

Ciascuna di queste teste è stata addestrata separatamente, al fine di ottimizzare le prestazioni su ciascun sottocompito; i risultati ottenuti nei rispettivi processi di addestramento sono presentati nella tabella 4.3.

Tabella 4.3: Metriche di test per il modello *all-roberta-large-v1*, includendo i risultati delle emozioni primarie e secondarie.

Tipo di Test	accuracy	Precisione	F1 Score	Recall
Primary Head	0.8369	0.8376	0.8366	0.8369
Secondary_Head_Anger	0.8768	0.8768	0.8758	0.8768
Secondary_Head_Anticipation	0.9160	0.9161	0.9151	0.9160
Secondary_Head_Disgust	0.9692	0.9697	0.9693	0.9692
Secondary_Head_Fear	0.7843	0.7844	0.7819	0.7843
Secondary_Head_Joy	0.9216	0.9214	0.9212	0.9216
Secondary_Head_Surprise	0.8291	0.8315	0.8289	0.8291
Secondary_Head_Trust	0.8627	0.8632	0.8628	0.8627
Secondary_Head_Sadness	0.7871	0.7851	0.7851	0.7871

Per queste teste, vengono mostrate anche le matrici di confusione.

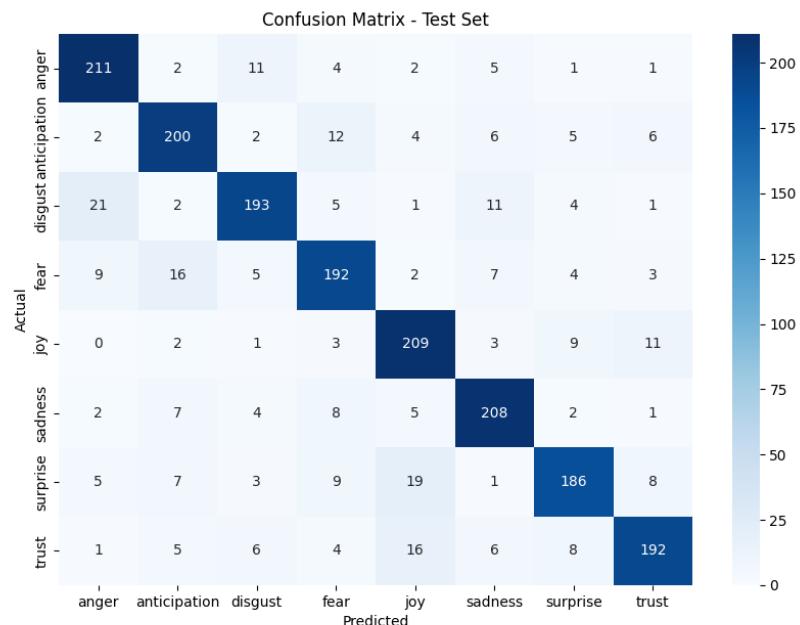


Figura 4.10: Matrice di confusione relativa alla testa primaria

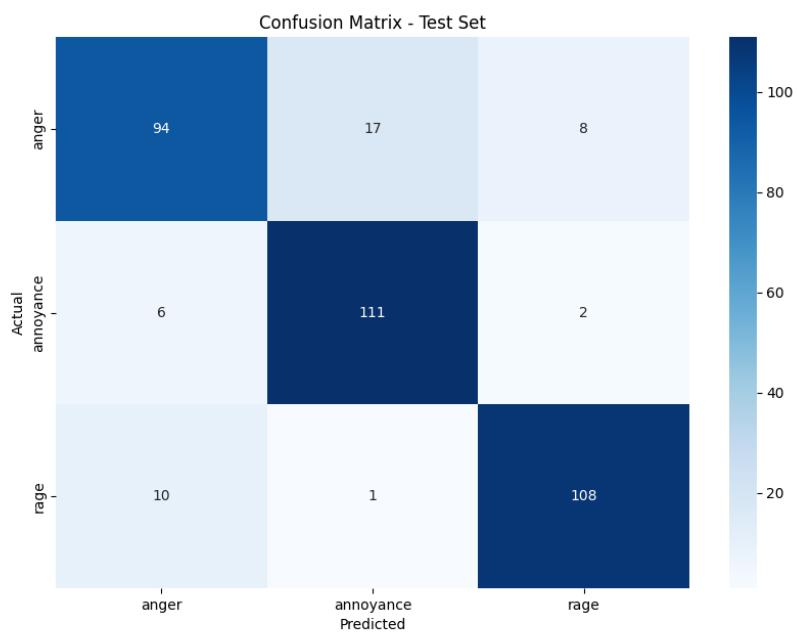


Figura 4.11: Matrice di confusione della testa secondaria per rabbia

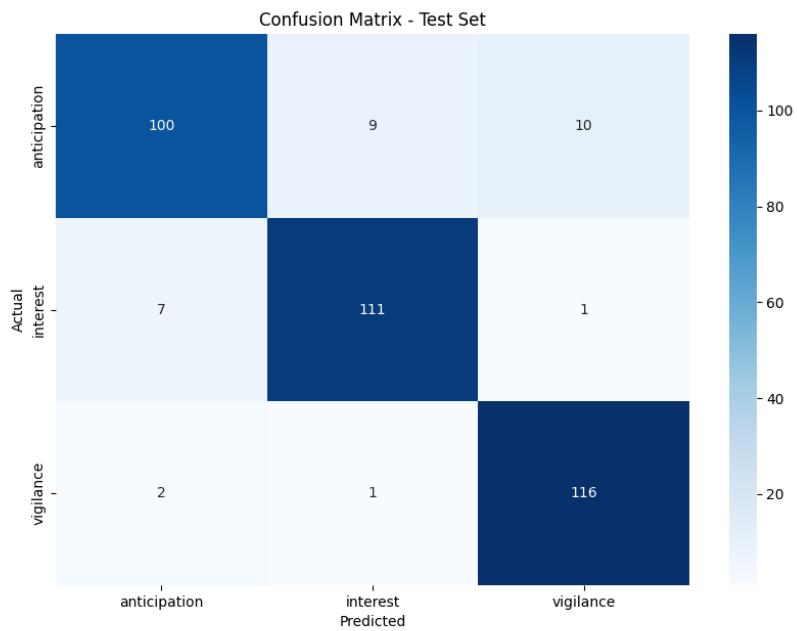


Figura 4.12: Matrice di confusione della testa secondaria per anticipazione

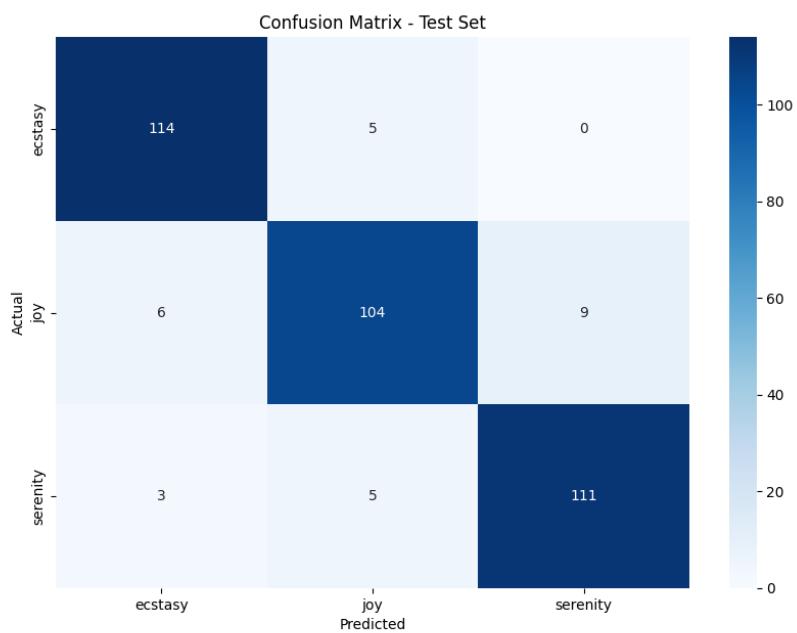


Figura 4.13: Matrice di confusione della testa secondaria per gioia

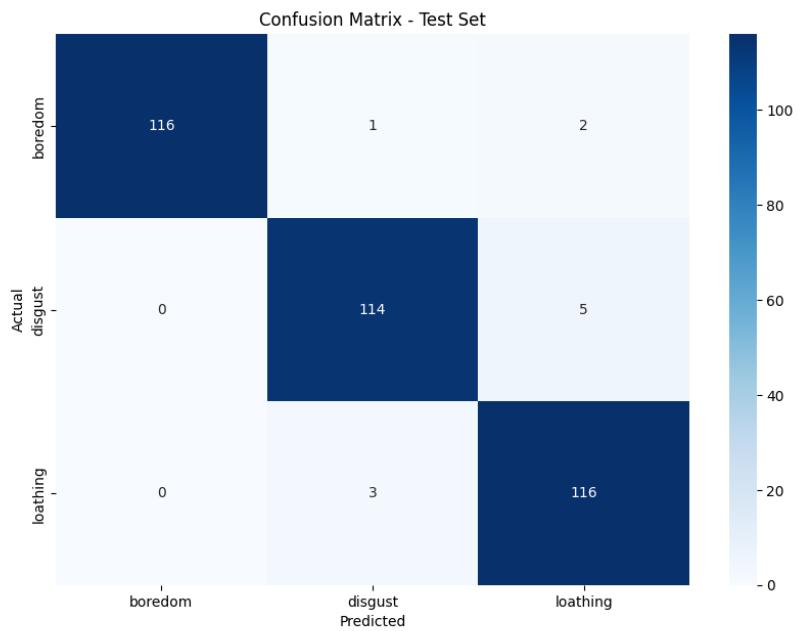


Figura 4.14: Matrice di confusione della testa secondaria per disgusto

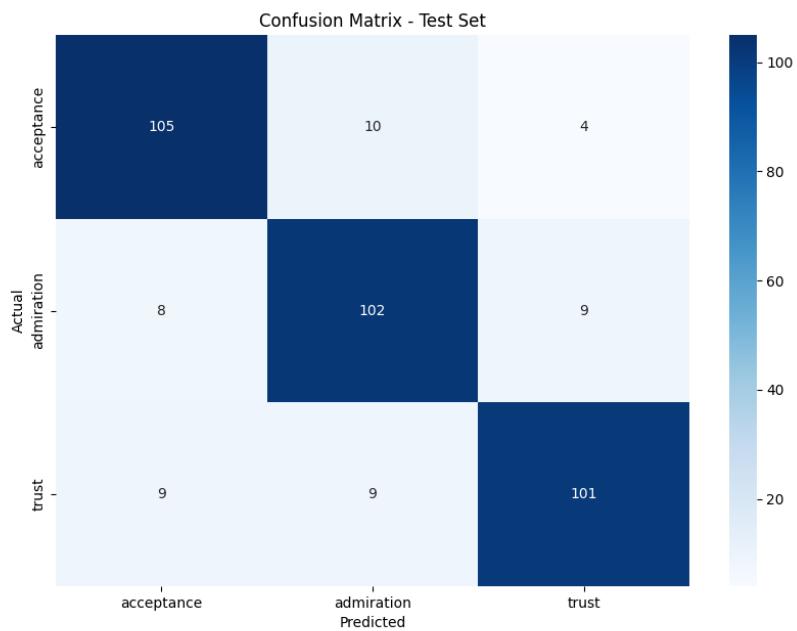


Figura 4.15: Matrice di confusione della testa secondaria per fiducia

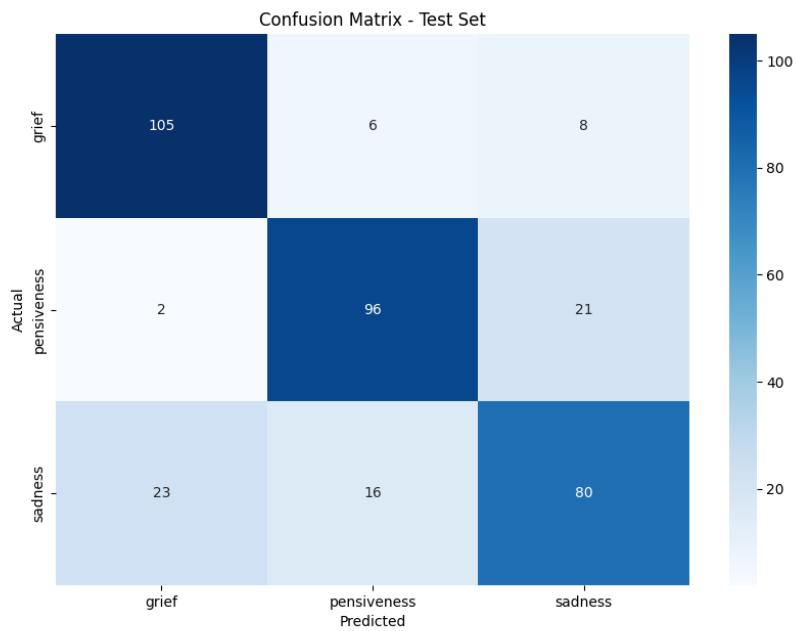


Figura 4.16: Matrice di confusione della testa secondaria per tristezza

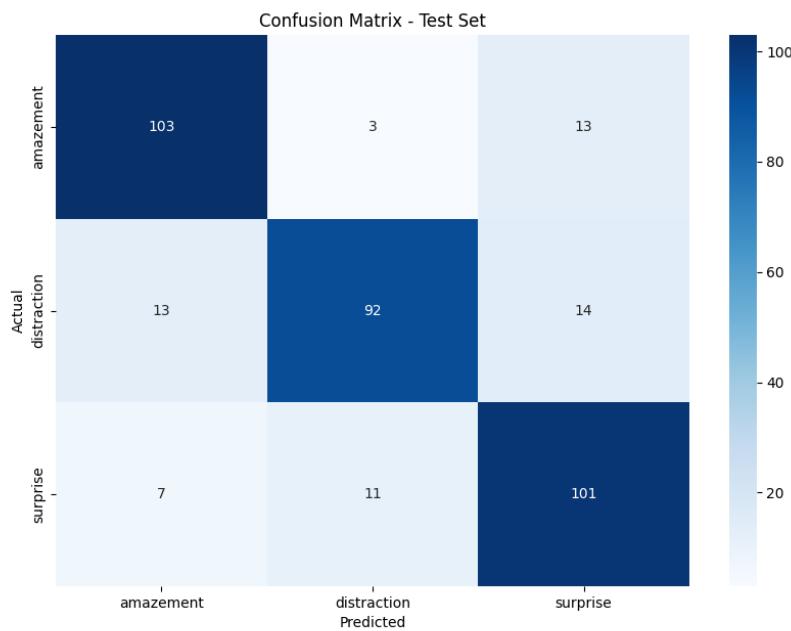


Figura 4.17: Matrice di confusione della testa secondaria per sorpresa

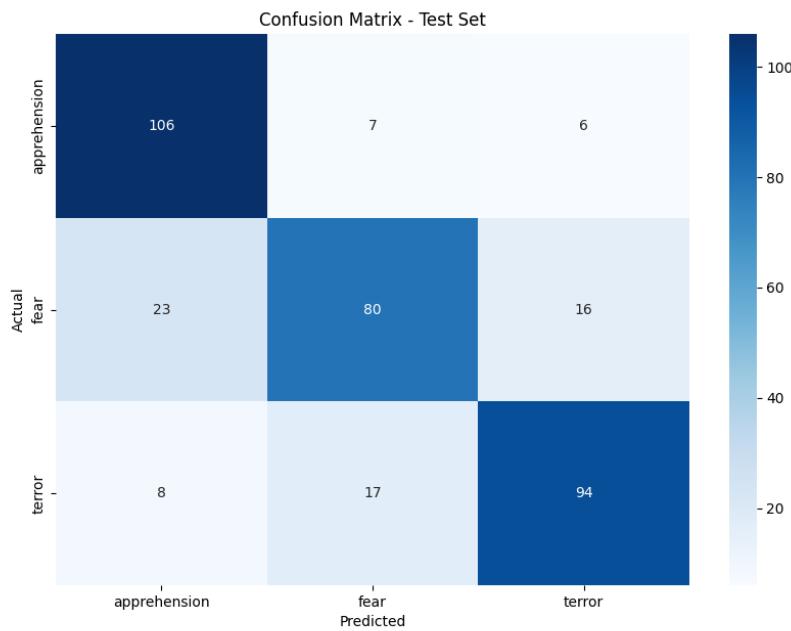


Figura 4.18: Matrice di confusione della testa secondaria per paura

L'analisi delle matrici di confusione relative alla testa primaria e alle teste secondarie evidenzia che il modello è in grado di classificare con buona

accuracy le emozioni previste, raggiungendo prestazioni elevate.

Gli errori residui risultano essere principalmente concentrati in corrispondenza di etichette emozionali facenti parte della stessa metaclasse, come da definizione nella sezione 3.3.

In particolare, dalla matrice di confusione della testa primaria emerge una leggera sovrapposizione tra emozioni come *trust* e *joy*, oppure tra *disgust* e *anger*, cioè coppie di emozioni vicine in Plutchick. Questo fenomeno suggerisce che il modello ha appreso rappresentazioni robuste in grado di distinguere efficacemente tra emozioni distinte, ma che in alcuni casi tende a confondere quelle che si collocano in regioni contigue nello spazio affettivo.

Anche tra le teste secondarie (v sez. 3.3), si osserva che gli errori si concentrano prevalentemente tra livelli di intensità adiacenti, difficilmente separabili poiché per definizione variazioni della stessa emozione, di natura sfumata e continua. Tali errori sono pertanto da considerarsi coerenti con il contesto applicativo, in quanto riflettono la complessità intrinseca del fenomeno emozionale nel modello di Plutchick, in cui la classificazione esatta nella dimensione dell'arousal resta dunque un problema aperto.

Per approfondire ulteriormente l'analisi delle misclassification, vediamo i risultati della classificazione delle emozioni mappate nelle tre metaclassi basate sul sentimento: Positive, Negative e Neutral (v. sez. 3.3); questa mappatura ci permette di verificare se gli errori di classificazione si verificano all'interno della stessa metaclasse oppure tra metaclassi differenti, ad un livello di astrazione maggiore di quello proposto dalla struttura di vicinanza dei petali di Plutchick.

Tabella 4.4: accuracy per le metaclassi (positive, neutral, negative) su *emotion_label2* (emozioni primarie) e *emotion_label1* (emozioni secondarie) utilizzando il modello *all-roberta-large-v1*.

Metaclasse	accuracy
Positive (Primarie)	0.9767
Neutral (Primarie)	0.9483
Negative (Primarie)	0.9792
Positive (Secondarie)	1.0000
Neutral (Secondarie)	0.8097
Negative (Secondarie)	0.9882

La Tabella 4.4 presenta le performance del modello *all-roberta-large-v1* nella classificazione gerarchica delle emozioni, suddivisa tra emozioni primarie (*emotion_label2*) e secondarie (*emotion_label1*), raggruppate secondo le metaclassi Positive, Neutral e Negative.

Per quanto riguarda le emozioni primarie, il modello dimostra un'elevata capacità discriminativa, con valori di accuratezza superiori al 94% su tutte le metaclassi. La performance migliore si osserva per la metaclasse Negative (97.92%), seguita da Positive (97.67%) e infine da Neutral (94.83%), indicando comunque una solida generalizzazione su tutte le categorie.

Nella classificazione delle emozioni secondarie si conferma un'elevata accuratezza complessiva, con la metaclasse Positive che raggiunge il massimo livello di performance (100%), seguita da Negative (98.82%). Tuttavia, si rileva una diminuzione dell'accuratezza per la metaclasse Neutral (80.97%).

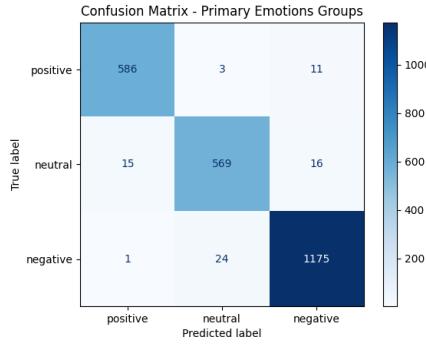


Figura 4.19: Matrice di confusione relativa alla testa primaria per le 3 metaclassi

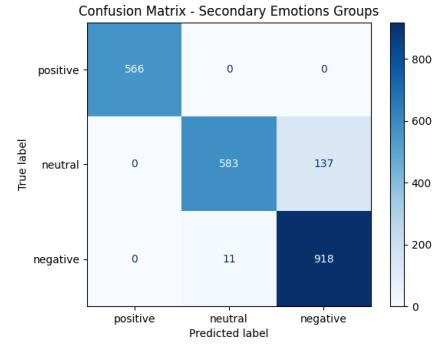


Figura 4.20: Matrice di confusione della testa secondaria per le 3 metaclassi

Le matrici di confusione 4.19 e 4.20 offrono un’analisi dettagliata delle prestazioni del modello *all-roberta-large-v1* nella classificazione delle metaclassi relative alle emozioni primarie (*emotion_label2*) e secondarie (*emotion_label1*).

Per le emozioni primarie, il modello dimostra un’elevata accuratezza, con un totale di 1976 esempi. La classe Positive (598 esempi) presenta una precisione del 97,66% (584/598 corretti), con soli 3 errori verso Neutral e 11 verso Negative. La classe Neutral (590 esempi) raggiunge il 96,61% di accuratezza (570/590 corretti), con 15 errori verso Positive e 16 verso Negative. La classe Negative (788 esempi) ottiene il 99,11% di accuratezza (781/788 corretti), con 1 errore verso Positive e 24 verso Neutral. Questi risultati, con un’accuratezza complessiva superiore al 97%, confermano l’efficacia del modello nel distinguere le principali categorie affettive.

Per le emozioni secondarie, su un totale di 1772 esempi, la classe Positive (566 esempi) mostra una precisione del 100% (566/566 corretti), senza errori. La classe Neutral (583 esempi) raggiunge il 91,25% di accuratezza (532/583 corretti), con 51 errori verso Negative. La classe Negative (623 esempi) ottiene il 98,56% di accuratezza (614/623 corretti), con 9 errori verso Neutral. Questi dati evidenziano una classificazione eccellente per

Positive, ma una maggiore incertezza nella distinzione tra Neutral e Negative, con un'accuratezza per Neutral del 91,25%, suggerendo un margine di miglioramento per questa metaclasse.

Il modello, quindi, si distingue per l'elevata efficacia nella classificazione delle emozioni primarie e delle emozioni secondarie Positive, mentre mostra una discriminazione meno netta tra le categorie Neutral e Negative nelle emozioni secondarie, con un'accuratezza complessiva che riflette la necessità di ulteriori ottimizzazioni in tale ambito.

La valutazione delle teste di classificazione è stata condotta in due fasi distinte, entrambe basate su dati mai presentati al modello durante la fase di addestramento:

1. Valutazione sul test set derivante dallo splitting dinamico:

ottenuto mediante la suddivisione automatica del *data set 1*, utilizzato durante la fase di addestramento come descritto nella Sezione 2.5. Il dettaglio della procedura di suddivisione nei sottoinsiemi di *training*, *validation* e *test set* è riportato nella sottosezione 3.2. In questa fase, l'analisi delle prestazioni si è concentrata esclusivamente sul test set così ottenuto.

2. Valutazione sul data set parte 2: costituito da un insieme separato e indipendente, mantenuto integralmente escluso dal processo di addestramento, al fine di disporre di un insieme ampio e rappresentativo di dati su cui effettuare una valutazione oggettiva, robusta e generalizzabile delle prestazioni del sistema. Questo data set, è descritto nella sezione 2.5

Tabella 4.5: Metriche di performance del modello *all-roberta-large-v1* su frasi mai viste prima, includendo emozioni primarie e secondarie.

Tipo di Test	accuracy	Precisione	F1 Score	Recall
Emozione Primaria	0.9229	0.9267	0.9223	0.9229
Emozione Secondaria (solo primaria corretta)	0.6772	0.7291	0.6609	0.6722

Le metriche mostrano che il modello, formato dalle teste descritte in precedenza organizzate nella struttura gerarchica definita nel capitolo precedente, ha un'elevata accuracy (92.29%) per le emozioni primarie, ma una performance inferiore (67.72%) per le secondarie. La bontà complessiva del modello gerarchico nel predire emozioni primarie e secondarie corrette è del 62.50% su 2400 frasi di test, considerando solo i casi in cui predice correttamente sia l'etichetta primaria che la secondaria.

4.3.1 Risultati Classificazione Stati Affettivi Composti

In tabella 4.6 confrontiamo l'accuracy del modello *all-roberta-large-v1* nel rilevare stati affettivi composti in frasi di test, come descritto nella sezione 3.4.

Per individuare ciascuno dei due flussi, indicheremo il primo approccio con *Sim(Similarity_Threshold)* e il secondo con *Sim/Prob(Similarity_Threshold e Probability_Threshold)*, descritte in dettaglio nella sezione 3.4.

L'approccio *Sim* mostra accuracy superiori, mentre l'aggiunta del controllo di probabilità in *Sim/Prob* vediamo accuracy ridotta, soprattutto per soglie di confidenza relativa più alte.

Tabella 4.6: Confronto dell’accuracy del modello *all-roberta-large-v1* per il rilevamento di stati affettivi complessi, con soglia di similarità (*Sim*) e con soglia di similarità e confidenza relativa dinamica (*Sim/Prob*).

Soglia (Sim/Prob)	accuracy Sim	accuracy <i>Sim/Prob</i>
0.60/–	0.5025	–
0.60/0.01	–	0.3563
0.60/0.05	–	0.2300
0.60/0.10	–	0.1213
0.65/–	0.5025	–
0.65/0.01	–	0.3563
0.65/0.05	–	0.2300
0.65/0.10	–	0.1213
0.70/–	0.5000	–
0.70/0.01	–	0.3550
0.70/0.05	–	0.2300
0.70/0.10	–	0.1213
0.75/–	0.4000	–
0.75/0.01	–	0.2725
0.75/0.05	–	0.1813
0.75/0.10	–	0.0925

Per la configurazione con sola soglia di similarità, l’accuracy rimane relativamente stabile per valori di *Sim* compresi tra 0.60 e 0.70, con valori intorno a 0.50 (0.5025 per *Sim*=0.60 e 0.65, 0.5000 per *Sim*=0.70), ma diminuisce significativamente a 0.4000 per *Sim*=0.75, suggerendo che una soglia di similarità troppo elevata varia il numero di emozioni composte rilevate, aumentando i falsi negativi.

Nella configurazione combinata *Sim/Prob*, l’accuracy decresce all’aumentare della soglia di confidenza relativa (*Prob*): per *Sim*=0.60, si passa da 0.3563

($\text{Prob}=0.01$) a 0.1213 ($\text{Prob}=0.10$), con un trend analogo per $\text{Sim}=0.65$ e $\text{Sim}=0.70$. Per $\text{Sim}=0.75$, l'accuracy è ancora più bassa, raggiungendo un minimo di 0.0925 con $\text{Prob}=0.10$.

Questo indica che l'introduzione della soglia di confidenza relativa, specialmente a valori più alti, rende il modello più selettivo, riducendo il richiamo e limitando il numero di emozioni composte identificate.

Complessivamente, la configurazione con sola soglia di similarità a 0.60 o 0.65 offre il miglior bilanciamento tra precisione e capacità di rilevamento, mentre l'approccio combinato Sim/Prob tende a penalizzare le prestazioni, particolarmente con soglie di probabilità più restrittive; da tale valutazione si evince che la miglior soluzione per questo task specifico, è utilizzare il metodo Sim con una soglia di similarità di 0.65.

Il sistema finale integra i componenti che hanno mostrato le migliori prestazioni negli esperimenti 3 e 4, combinando un approccio ibrido alla classificazione delle emozioni. In particolare, vengono utilizzate le teste di classificazione fine-tunate sul modello RoBERTa, specializzate rispettivamente nella rilevazione delle emozioni primarie e secondarie, affiancate da un modulo di valutazione semantica delle etichette emozionali. Quest'ultimo sfrutta la similarità semantica calcolata mediante *FAISS* (Facebook AI Similarity Search) e le rappresentazioni vettoriali generate dal modello *Sentence-T5*, al fine di identificare con maggiore precisione le emozioni composte. L'architettura completa è visibile in figura 3.3.

4.4 Rilevamento Emozioni nel Grooming

Dopo aver validato il sistema migliore per il task del rilevamento di emozioni, lo abbiamo applicato al grooming.

In questa sezione riportiamo i risultati ottenuti dall'applicazione del siste-

ma di classificazione al data set di grooming dataset_sintetico_grooming presentato nella sezione 2.5.

L'analisi delle emozioni rilevate evidenzia pattern significativi associati alle diverse fasi (v. sez. 1.1) del processo di adescamento, fornendo un primo riscontro empirico sulla distribuzione emotionale nelle fasi di grooming.

Tabella 4.7: Conteggi Assoluti delle **Emozioni Primarie** per Tag

Tag	Total	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Compliment	351	0	24	0	1	193	10	6	117
Relationship	330	0	24	0	1	186	10	6	103
Activities	316	0	88	1	7	157	30	17	16
Approach	308	9	45	19	58	42	42	75	18
Communicative Desensitization	358	3	67	22	40	68	81	57	20
Personal Information	357	5	45	1	27	111	60	23	85
Reframing	194	2	34	2	15	39	34	22	46
Isolation	365	10	25	1	79	20	143	2	85

Tabella 4.8: Conteggi Assoluti delle **Emozioni Primarie** per Fase

Phase	Total	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Fulfilling Needs	351	0	24	0	1	193	10	6	117
Isolation	917	15	126	25	134	127	258	81	151
Sexualizing Relationship	308	9	45	19	58	42	42	75	18
Targeting and Gaining Trust	1003	5	157	2	35	454	100	46	204

Nella fase di *Targeting and Gaining Trust* (*Individuazione della vittima e acquisizione della fiducia*), l'emozione predominante è *joy* (*Gioia*), con 454 occorrenze (45,27% del totale rispetto alla singola fase) seguita da *trust* (*Fiducia*) (20,34%) e *anticipation* (*Anticipazione*) (15,65%), denotando la prenza di emozioni riconducibili alla metaclasse emotionale Positiva.

La fase di *Isolation* (*Isolamento*) mostra un cambiamento netto nel profilo emotionale: aumenta significativamente la presenza di *sadness* (*Tristezza*) (258 occorrenze), seguita da *fear* (*Paura*) (134) e *trust* (*Fiducia*) (151). Durante la fase *Sexualizing the Relationship* (*Sessualizzazione della rela-*

zione), emerge in modo prevalente *surprise* (*Sorpresa*) (75 occorrenze), accompagnata da *fear* (58), *anticipation* (45), e *sadness* (42).

Nella fase di *Fulfilling Needs* (*Soddisfacimento dei bisogni*), le emozioni ritornano ad assumere una connotazione più positiva, con una prevalenza di *joy* (193) e *trust* (117).

Tabella 4.9: Conteggi Assoluti delle **Emozioni Secondarie** per Tag

Tag	Total	Acceptance	Admiration	Amazement	Anger	Annoyance	Anticipation	Apprehension	Boredom	Disgust	Distraction	Ecstasy	Fear	Grief	Interest	Joy	Loathing	Pensiveness	Rage	Sadness	Serenity	Surprise	Terror	Trust	Vigilance
Activities	316	10	2	12	0	0	41	4	0	0	5	1	1	7	44	149	1	15	0	8	7	0	2	4	3
Approach	308	14	0	58	3	0	37	24	4	15	15	2	8	22	7	40	0	10	6	10	0	2	26	4	1
Communicative Desensitization	358	18	0	37	0	2	44	25	4	13	15	1	5	24	14	64	5	41	1	16	3	5	10	2	9
Compliment	351	34	44	3	0	0	1	1	0	0	3	2	0	0	13	176	0	5	0	5	15	0	0	39	10
Isolation	365	53	0	1	6	0	20	57	1	0	1	0	19	100	1	20	0	18	4	25	0	0	3	32	4
Personal Information	357	62	4	12	0	1	27	18	1	0	7	0	6	3	14	104	0	50	4	7	7	4	3	19	4
Reframing	194	33	6	18	2	0	18	5	2	0	4	0	5	19	16	39	0	11	0	4	0	0	5	7	0
Relationship	330	32	40	3	0	0	1	1	0	0	3	2	0	0	13	169	0	5	0	5	15	0	0	31	10

Tabella 4.10: Conteggi Assoluti delle **Emozioni Secondarie** per Fase

Phase	Total	Acceptance	Admiration	Amazement	Anger	Annoyance	Anticipation	Apprehension	Boredom	Disgust	Distraction	Ecstasy	Fear	Grief	Interest	Joy	Loathing	Pensiveness	Rage	Sadness	Serenity	Surprise	Terror	Trust	Vigilance
Fulfilling Needs	351	34	44	3	0	0	1	1	0	0	3	2	0	0	13	176	0	5	0	5	15	0	0	39	10
Isolation	917	104	6	56	8	2	82	87	7	13	20	1	29	143	31	123	5	70	5	45	3	5	18	41	13
Sexualizing Relationship	308	14	0	58	3	0	37	24	4	15	15	2	8	22	7	40	0	10	6	10	0	2	26	4	1
Targeting and Gaining Trust	1003	104	46	27	0	1	69	23	1	0	15	3	7	10	71	422	1	70	4	20	29	4	5	54	17

L'analisi delle emozioni secondarie conferma le stesse dinamiche: *joy* è ancora dominante in *Targeting and Gaining Trust* (422 occorrenze), mentre *grief* (*Lutto*) emerge con forza in *Isolation* (143). Tra gli stati affettivi composti, si osservano elevati livelli di *optimism* (*Ottimismo*) (425) nella fase iniziale, di *love* (*Amore*) (201) in *Fulfilling Needs* e allo stesso modo, l'emozione di *submission* (*Sottomissione*) (154), particolarmente presente nella fase di *Isolation*.

Tabella 4.11: Conteggi Assoluti delle **Stati affettivi** per **Tag**

Tag	Total	Aggressiveness	Awe	Contempt	Disapproval	Love	None	Optimism	Remorse	Submission
Activities	316	2	7	0	33	49	17	200	5	5
Approach	308	8	87	8	26	21	24	64	15	19
Communicative Desensitization	358	5	48	5	53	39	45	83	23	15
Compliment	351	0	1	0	3	201	13	131	0	2
Isolation	365	7	11	3	25	73	92	12	24	118
Personal Information	357	6	14	3	25	104	59	100	5	31
Reframing	194	0	18	5	21	38	29	53	7	21
Relationship	330	0	1	0	3	186	13	125	0	2

Tabella 4.12: Conteggi Assoluti delle **Stati affettivi** per **Fase**

Phase	Total	Aggressiveness	Awe	Contempt	Disapproval	Love	None	Optimism	Remorse	Submission
Fulfilling Needs	351	0	1	0	3	201	13	131	0	2
Isolation	917	12	77	13	99	150	166	148	54	154
Sexualizing Relationship	308	8	87	8	26	21	24	64	15	19
Targeting and Gaining Trust	1003	8	22	3	61	339	89	425	10	38

Tra gli stati affettivi composti, si osservano elevati livelli di *optimism* (425 occorrenze) nella fase di *Targeting and Gaining Trust*, mentre nella fase *Fulfilling Needs*, lo stato affettivo di *love* (201) è dominante.

Nella fase di *Isolation*, emerge invece un'alta presenza di *submission* (154), accompagnata da *remorse* (*Rimorso*) (54) e *disapproval* (*Disapprovazione*) (99).

Nella fase *Sexualizing the Relationship*, lo stato affettivo di *awe* (*Stupore*) (87) è prevalente, seguito da *contempt* (*Disprezzo*) (8) e *aggressiveness* (*Aggressività*) (8). La categoria *None* (*Nessuno stato affettivo*), con 166 occorrenze in *Isolation*, 89 in *Targeting and Gaining Trust*, 24 in *Sexualizing the Relationship* e 13 in *Fulfilling Needs*, cattura le interazioni in cui non emergono stati affettivi composti chiaramente definiti, che potrebbero essere collegati a risposte ambigue, neutre o complesse. Si evidenzia dunque in questa fase le difficoltà che emergono in alcuni casi, nel classificare in termini di stati affettivi composti le reazioni della vittima in contesti manipolatori.

Da questa panoramica, i dati raccolti suggeriscono un'evoluzione dinamica e una differenziazione del profilo emozionale nel corso delle diverse fasi/tag.

Analizziamo ora più in dettaglio i risultati, inserendo al termine di ogni sezione degli histogrammi riassuntivi, per evidenziare i pattern salienti e le informazioni significative principali. Nei grafici, le emozioni primarie saranno rappresentate in blu, le secondarie in arancione e stati affettivi composti in verde, fornendo una vista d'insieme.

Frequenze Relative

Le frequenze relative delle emozioni primarie, secondarie e degli stati affettivi, descritte nella sezione 3.5, risultano fondamentali in quanto permettono di rendere i valori confrontabili, indipendentemente dalla numerosità di ciascuna categoria emotiva. Le tabelle 4.13, 4.14, 4.15, 4.16, 4.17, 4.18 normalizzano la frequenza assoluta delle emozioni rilevate in ciascuna fase (o tag) del processo, restituendo valori in percentuale.

1. Emozioni Primarie per Tag

Tabella 4.13: Frequenze relative delle emozioni primarie per tag

Tag	Emozioni Principali	Frequenza (%)
Activities (Totale = 316)	Joy, Anticipation	49.7%, 27.8%
Approach (Totale = 308)	Surprise, Fear	24.4%, 18.8%
Communicative Desensitization (Totale = 358)	Sadness, Joy	22.6%, 19.0%
Compliment (Totale = 351)	Joy, Trust	55.0%, 33.3%
Isolation (Totale = 365)	Sadness, Trust	39.2%, 23.3%
Personal Information (Totale = 357)	Joy, Trust	31.1%, 23.8%
Reframing (Totale = 194)	Trust, Joy	23.7%, 20.1%
Relationship (Totale = 330)	Joy, Trust	56.4%, 31.2%

Tra le emozioni primarie, *Joy* domina nei tag *Compliment* (55.0%), *Relationship* (56.4%), *Activities* (49.7%) e *Personal Information* (31.1%); *Trust* è rilevante in *Compliment* (33.3%), *Relationship* (31.2%), *Isolation* (23.3%) e *Personal Information* (31.1%). *Sadness* prevale in *Isolation* (39.2%) mentre

Surprise è corretalato ad *Approach* (24.4%).

2. Emozioni Primarie per Fase

Tabella 4.14: Frequenze relative delle emozioni primarie per fase

Phase	Emozioni Principali	Frequenza (%)
Fulfilling Needs (Totale = 351)	Joy, Trust	55.0%, 33.3%
Isolation (Totale = 917)	Sadness, Trust	28.1%, 16.5%
Sexualizing Relationship (Totale = 308)	Surprise, Fear	24.4%, 18.8%
Targeting and Gaining Trust (Totale = 1003)	Joy, Trust	45.2%, 20.3%

Nella fase *Fulfilling Needs*, le emozioni principali sono *Joy* e *Trust*, con frequenze rispettivamente del 55,0% e del 33,3%. Durante la fase di *Isolation*, emerge un cambiamento marcato con *Sadness* al 28,1% e *Trust* al 16,5%. La fase di *Sexualizing Relationship* vede invece prevalere emozioni come *Surprise* al 24,4% e *Fear* al 18,8%. Infine, nella fase di *Targeting and Gaining Trust*, tornano preponderanti *Joy* al 45,2% e *Trust* al 20,3%.

3. Emozioni Secondarie per Tag

Tabella 4.15: Frequenze relative delle emozioni secondarie per tag

Tag	Emozioni	Frequenza (%)
Activities (Totale = 316)	Joy, Interest	47.2%, 13.9%
Approach (Totale = 308)	Amazement, Anticipation	18.8%, 12.0%
Communicative Desensitization (Totale = 358)	Joy, Pensiveness	17.9%, 11.5%
Compliment (Totale = 351)	Joy, Admiration	50.1%, 12.5%
Isolation (Totale = 365)	Grief, Acceptance	27.4%, 14.5%
Personal Information (Totale = 357)	Joy, Acceptance	29.1%, 17.4%
Reframing (Totale = 194)	Joy, Acceptance	20.1%, 17.0%
Relationship (Totale = 330)	Joy, Admiration	51.2%, 12.1%

4. Emozioni Secondarie per Fase

Tabella 4.16: Frequenze relative delle emozioni secondarie per fase

Phase	Emozioni	Frequenza (%)
Fulfilling Needs (Totale = 351)	Joy, Admiration	50.1%, 12.5%
Isolation (Totale = 917)	Grief, Joy	15.6%, 13.4%
Sexualizing Relationship (Totale = 308)	Amazement, Anticipation	18.8%, 12.0%
Targeting and Gaining Trust (Totale = 1003)	Joy, Acceptance	42.1%, 10.4%

Tra le emozioni secondarie, *Joy* è prominente in *Compliment* (50.1%) e *Relationship* (51.2%), mentre *Grief* emerge in *Isolation* (27.4% per tag, 15.6% per fase). *Amazement* è rilevante in *Approach* (18.8%) e nella fase di *Sexualizing Relationship* (18.8%).

5. Stati Affettivi Composti per Tag

Tabella 4.17: Frequenze relative degli Stati Affettivi Composti per tag

Tag	Stati Affettivi	Frequenza (%)
Activities (Totale = 316)	Optimism, Love	63.3%, 15.5%
Approach (Totale = 308)	Awe, Optimism	28.2%, 20.8%
Communicative Desensitization (Totale = 358)	none, Optimism	24.3%, 23.2%
Compliment (Totale = 351)	Love, Optimism	57.3%, 37.3%
Isolation (Totale = 365)	Submission, none	32.3%, 25.2%
Personal Information (Totale = 357)	Love, Optimism	29.1%, 28.0%
Reframing (Totale = 194)	Optimism, Love	27.3%, 19.6%
Relationship (Totale = 330)	Love, Optimism	56.4%, 37.9%

6. Stati Affettivi Composti per Fase

Tabella 4.18: Frequenze relative degli Stati Affettivi Composti per fase

Phase	Stati Affettivi	Frequenza (%)
Fulfilling Needs (Totale = 351)	Love, Optimism	57.3%, 37.3%
Isolation (Totale = 917)	none, Submission	22.9%, 16.8%
Sexualizing Relationship (Totale = 308)	Awe, Optimism	28.2%, 20.8%
Targeting and Gaining Trust (Totale = 1003)	Optimism, Love	42.4%, 33.8%

Tra gli Stati Affettivi Composti, *Optimism* domina nel tag *Activities* (63.3%) e nella fase *Targeting and Gaining Trust* (42.4%), mentre *Love* prevale nei tag *Compliment* (57.3%) e *Relationship* (56.4%). *Submission* è rilevante in *Isolation* (32.3% per tag, 16.8% per fase).

None è prominente in *Communicative Desensitization* (24.3%) e *Isolation* (25.2% per tag, 22.9% per fase). Emozioni della metaclasse Negative, come *Anger* e *Disgust*, hanno incidenza minima (< 6.2%) e sono concentrate nel tag e nella fase *isolation*, indicando una preferenza per emozioni delle metaclassi Positive o Neutral, come confermato dai grafici 4.21 e 4.22, dove viene presentata una vista aggregata delle principali frequenze relative delle emozioni, associate ai tag e alle fasi corrispondenti. In entrambi i casi, gli histogrammi mostrano una prevalenza di emozioni Positive come *joy*, *love*, *optimism*.

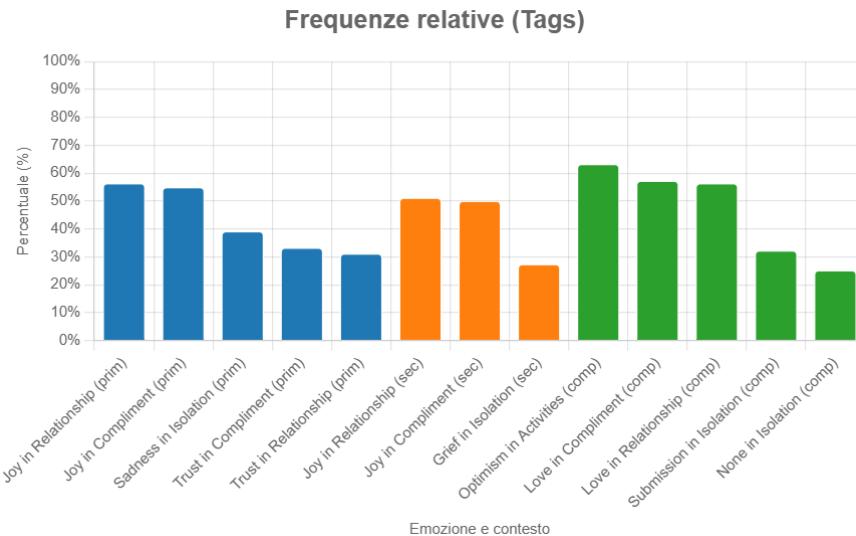


Figura 4.21: Frequenze relative rilevanti per Tags

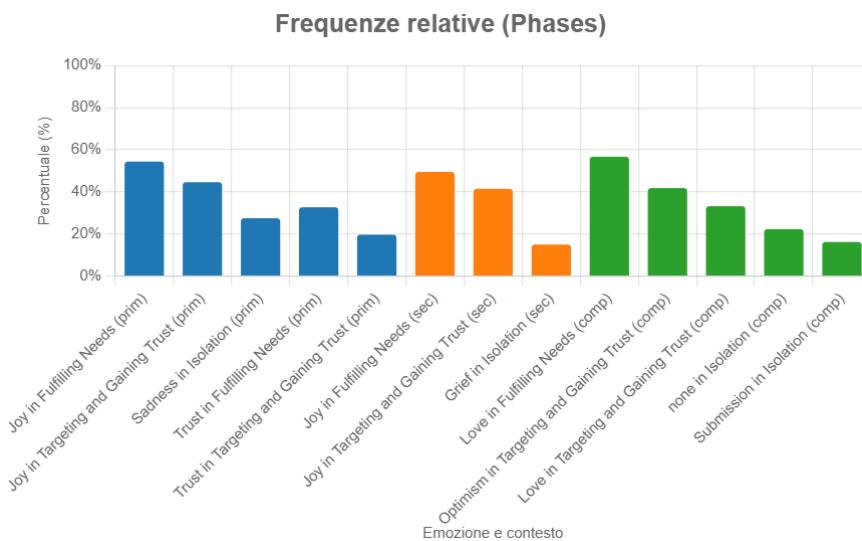


Figura 4.22: Frequenze relative rilevanti per Phases

Indice di Dominanza

Questo indice, spiegato nella sezione 3.5, evidenzia le emozioni predominanti in ciascun contesto. La scelta di tenere in considerazione sia la prima che la seconda emozione più frequente è dettata dalla volontà di ottenere una valutazione generalizzata e maggiormente rappresentativa relativamente alla distribuzione delle etichette per ciascuna fase/tag.

Tabella 4.19: Indice di Dominanza per Emozioni Primarie (Tag)

Tag	Emozioni Dominanti	Indice di Dominanza
Activities	Joy / Anticipation	0.4968 / 0.2785
Approach	Surprise / Anticipation	0.2435 / 0.1461
Communicative Desensitization	Sadness / Joy	0.2263 / 0.1899
Compliment	Joy / Trust	0.5499 / 0.3333
Isolation	Sadness / Trust	0.3918 / 0.2329
Personal Information	Joy / Trust	0.3109 / 0.2381
Reframing	Trust / Joy	0.2371 / 0.2010
Relationship	Joy / Trust	0.5636 / 0.3121

Tabella 4.20: Indice di Dominanza per Emozioni Primarie (Fasi)

Fase	Emozioni Dominanti	Indice di Dominanza
Fulfilling Needs	Joy / Trust	0.5499 / 0.3333
Isolation	Sadness / Trust	0.2814 / 0.1648
Sexualizing Relationship	Surprise / Anticipation	0.2435 / 0.1461
Targeting and Gaining Trust	Joy / Trust	0.4526 / 0.2036

Per le emozioni primarie, l'emozione *joy* domina nei tag *Compliment* ($D = 0.5499$, seconda: *trust*, $D = 0.3333$), *Relationship* ($D = 0.5636$, seconda: *trust*, $D = 0.3121$), *Activities* ($D = 0.4968$, seconda: *anticipation*, $D = 0.2785$), e nelle fasi *Fulfilling Needs* ($D = 0.5499$, seconda: *trust*, $D = 0.3333$) e *Targeting and Gaining Trust* ($D = 0.4526$, seconda: *trust*, $D = 0.2036$).

Il secondo indice di dominanza, quando assume valori significativamente elevati (ad esempio 0.3333 per il tag *Compliment*), evidenzia la presenza di una componente secondaria con un peso rilevante, indicando che l'attivazione emotiva non si limita all'emozione primaria, ma presenta una co-dominanza significativa della seconda emozione più presente. I tag *Approach* ($D = 0.2435$, *surprise*; seconda: *anticipation*, $D = 0.1461$), *Reframing* ($D = 0.2371$, *joy*; seconda: *trust*, $D = 0.2010$) e *Communicative Desensitization*

($D = 0.2263$, *sadness*; seconda: *joy*, $D = 0.1899$) e le fasi *Sexualizing Relationship* ($D = 0.2435$, *surprise*; seconda: *anticipation*, $D = 0.1461$) e *Isolation* ($D = 0.2814$, *sadness*; seconda: *trust*, $D = 0.1648$) mostrano indici di dominanza bassi per entrambe le emozioni, vicini alla distribuzione uniforme (0.125), indicando un'alta varietà emotiva.

I tag *Activities* ($D = 0.4968$, *joy*; seconda: *anticipation*, $D = 0.2785$), *Compliment* ($D = 0.5499$, *joy*; seconda: *trust*, $D = 0.3333$) e *Relationship* ($D = 0.5636$, *joy*; seconda: *trust*, $D = 0.3121$) mostrano una marcata differenza (maggiore di 0.20) tra i valori di dominanza della prima e della seconda emozione analizzata, suggerendo quindi forte dominanza della prima. Tra le fasi, invece, *Targeting and Gaining Trust* ($D = 0.4526$, *joy*; seconda: *trust*, $D = 0.2036$) e *Fulfilling Needs* ($D = 0.5499$, *joy*; seconda: *trust*, $D = 0.3333$), evidenziano una differenza di dominanza tra le due emozioni principali maggiore di 0.2. dall'analisi, quindi, si può dedurre che nella fasi *Targeting and Gaining Trust* e *Fulfilling Needs* e nei tag *activities*, *compliment*, *relationship* esiste una emozione predominante identificabile con *joy*. Negli altri casi, il divario tra prima e seconda emozione dominante è sottile e pertanto sono identificabili delle co-dominanze tra le due emozioni rilevate come principali.

Tabella 4.21: Indice di Dominanza per Emozioni Secondarie (Tag)

Tag	Emozioni Dominanti	Indice di Dominanza
Activities	Joy / Interest	0.4715 / 0.1392
Approach	Amazement / Anticipation	0.1883 / 0.1201
Communicative Desensitization	Joy / Anticipation	0.1788 / 0.1229
Compliment	Joy / Admiration	0.5014 / 0.1254
Isolation	Grief / Acceptance	0.2740 / 0.1452
Personal Information	Joy / Acceptance	0.2913 / 0.1737
Reframing	Joy / Acceptance	0.2010 / 0.1701
Relationship	Joy / Admiration	0.5121 / 0.1212

Tabella 4.22: Indice di Dominanza per Emozioni Secondarie (Fasi)

Fase	Emozioni Dominanti	Indice di Dominanza
Fulfilling Needs	Joy / Admiration	0.5014 / 0.1254
Isolation	Grief / Joy	0.1558 / 0.1341
Sexualizing Relationship	Amazement / Anticipation	0.1883 / 0.1201
Targeting and Gaining Trust	Joy / Acceptance	0.4207 / 0.1038

Per le emozioni secondarie, *joy* prevale con forte dominanza nei tag *Compliment* ($D = 0.5014$, seconda: *admiration*, $D = 0.1254$), *Relationship* ($D = 0.5121$, seconda: *admiration*, $D = 0.1212$), *Activities* ($D = 0.4715$, seconda: *interest*, $D = 0.1392$) e nelle fasi *Fulfilling Needs* ($D = 0.5014$, seconda: *admiration*, $D = 0.1254$) e *Targeting and Gaining Trust* ($D = 0.4207$, seconda: *acceptance*, $D = 0.1038$), evidenziando in modo ancora più marcato e confermando quanto già analizzato in precedenza, in riferimento alle tabelle 4.19 e 4.20.

I tag *Approach* ($D = 0.1883$, *amazement*; seconda: *anticipation*, $D = 0.1201$) e *Sexualizing Relationship* ($D = 0.1883$, *amazement*; seconda: *anticipation*, $D = 0.1201$) presentano indici di dominanza bassi, vicini alla distribuzione uniforme (0.0417), indicando una varietà di emozioni secondarie con simile numerosità.

Anche i tag *Communicative Desensitization* ($D = 0.1788$, *joy*; seconda: *anticipation*, $D = 0.1229$) e *Reframing* ($D = 0.2010$, *joy*; seconda: *acceptance*, $D = 0.1701$) mostrano indici di dominanza simili per entrambe le emozioni analizzate.

Tabella 4.23: Indice di Dominanza per Stati Affettivi Composti (Tag)

Tag	Stati Affettivi Composti	Indice di Dominanza
Activities	Optimism / Love	0.6329 / 0.1551
Approach	Awe / Optimism	0.2825 / 0.2078
Communicative Desensitization	None / Optimism	0.2430 / 0.2318
Compliment	Love / Optimism	0.5726 / 0.3732
Isolation	Submission / None	0.3233 / 0.2521
Personal Information	Love / Optimism	0.2913 / 0.2801
Reframing	Optimism / Love	0.2732 / 0.1959
Relationship	Love / Optimism	0.5636 / 0.3788

Tabella 4.24: Indice di Dominanza per Stati Affettivi Composti (Fasi)

Fase	Stati Affettivi Composti	Indice di Dominanza
Fulfilling Needs	Love / Optimism	0.5726 / 0.3732
Isolation	None / Submission	0.2289 / 0.1679
Sexualizing Relationship	Awe / Optimism	0.2825 / 0.2078
Targeting and Gaining Trust	Optimism / Love	0.4237 / 0.3377

L'analisi degli *Stati Affettivi Composti* per i tag evidenzia che *Activities* ($D = 0.6329$, *Optimism*; seconda: *Love*, $D = 0.1551$), *Compliment* ($D = 0.5726$, *Love*; seconda: *Optimism*, $D = 0.3732$) e *Relationship* ($D = 0.5636$, *Love*; seconda: *Optimism*, $D = 0.3788$), così come le fasi *Fulfilling Needs* ($D = 0.5726$, *Love*; seconda: *Optimism*, $D = 0.3732$) e *Targeting and Gaining* ($D = 0.4237$, *Optimism*; seconda: *Love*, $D = 0.3377$) mostrano elevati valori di dominanza per gli stati affettivi principali, con differenze marcate nei valori di dominanza tra la prima e la seconda emozione più frequente. Questa analisi evidenzia che, pur con valori numerici variabili dovuti al numero diverso di emozioni considerate tra primarie, secondarie e stati affettivi, l'indice di dominanza mostra che per i tag *Relationship*, *Activities* e per le fasi *Fulfilling Needs* e *Targeting and Gaining Trust*, la dominanza di una singola emozione, appartenente alla macroclasse Positive, è netta e marcata rispetto a tutte

le altre, come si deduce dalla differenza elevata riscontrabile dal confronto con l'indice della seconda emozione dominante. Indipendentemente dal tipo di emozione considerata (i.e., primaria, secondaria, stato affettivo), i valori ottenuti confermano questo andamento, mantendo sempre per le fasi e i tag appena definiti, i valori più alti rispetto a tutti gli altri, permettendo di definire questo pattern. Come si vede, nell'immagine 4.23 relativa ai tag, i valori più alti di dominanza sono presenti in corrispondenza di *Relationship*, *Compliment* e *Activities*, mentre nell'immagine 4.24 sono evidenti le fasi *Fulfilling Needs* e *Targeting and Gaining Trust* come quelle con valori più significativi.

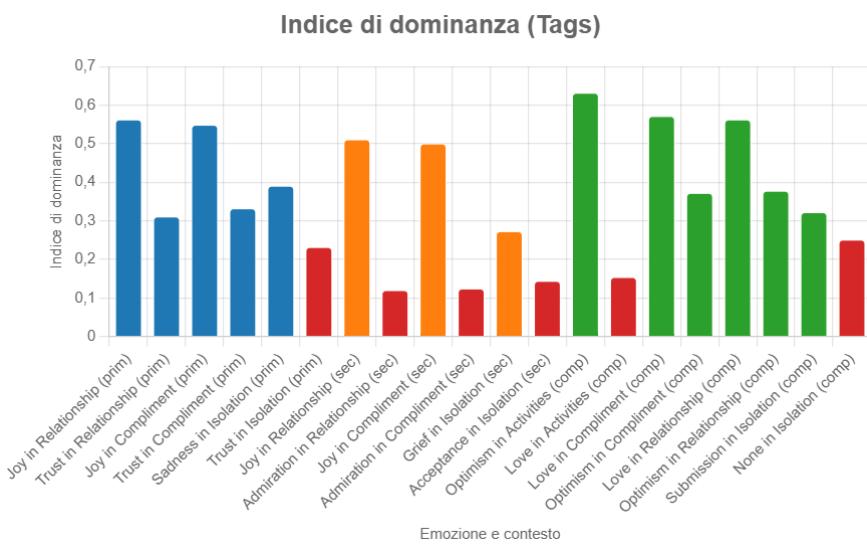


Figura 4.23: Indici di dominanza rilevanti per Tags

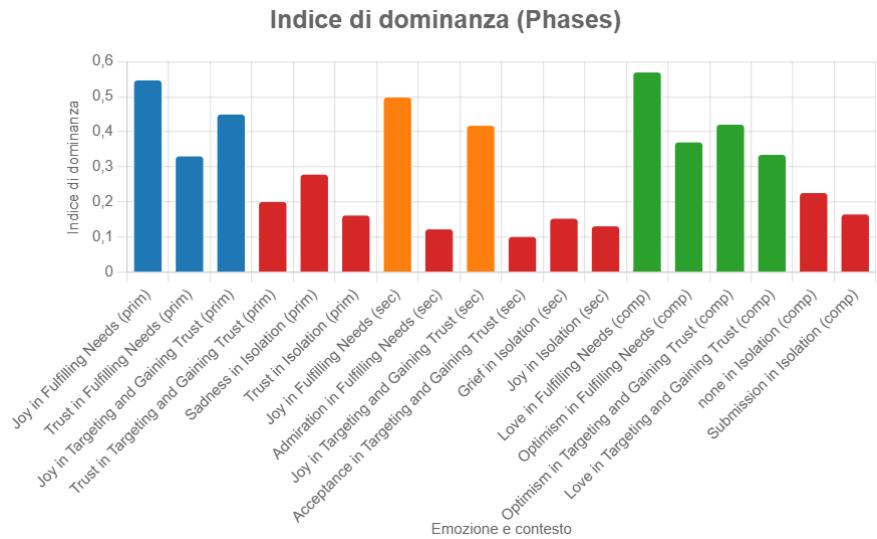


Figura 4.24: Indici di dominanza rilevanti per Phases

Entropia di Shannon

L'entropia di Shannon, invece, introdotto nella sezione 3.5, è utile per misurare l'incertezza di una distribuzione di probabilità. I valori di entropia sono espressi in bit e arrotondati a tre decimali per uniformità.

Tabella 4.25: Entropia di Shannon per Emozioni per Tag

Tag	Primarie (bit)	Secondarie (bit)	Stati Affettivi Composti (bit)
Activities	2.058	2.805	2.241
Approach	2.898	3.858	2.896
Communicative Desensitization	2.895	3.714	2.940
Compliment	1.797	2.726	1.858
Isolation	2.496	3.195	2.951
Personal Information	2.697	3.353	2.990
Reframing	2.848	3.574	2.904
Relationship	1.846	2.766	1.858

Tabella 4.26: Entropia di Shannon per Emozioni per Fasi

Fase	Primarie (bit)	Secondarie (bit)	Stati Affettivi Composti (bit)
Fulfilling Needs	1.797	2.726	1.858
Isolation	2.951	4.168	3.250
Sexualizing Relationship	2.898	3.858	2.896
Targeting and Gaining Trust	2.502	3.234	2.513

Il valore massimo teorico è $\log_2(n)$, dove n è il numero di emozioni possibili: $\log_2(8) = 3$ bit per le emozioni primarie, $\log_2(24) \approx 4.585$ bit per le secondarie e $\log_2(9) \approx 3.170$ bit per le composte.

Per le emozioni primarie, i tag *Compliment* ($H = 1.797$) e *Relationship* ($H = 1.846$) mostrano bassa entropia mentre, al contrario, i tag *Communicative Desensitization* ($H = 2.895$) e *Approach* ($H = 2.898$) presentano valori elevati.

Tra le fasi, *Fulfilling Needs* ($H = 1.797$) ottiene la più bassa entropia mentre *Isolation* ($H = 2.951$) si avvicina al massimo teorico di 3 bit, costituendo la fase con variabilità emotiva maggiore.

Per le emozioni secondarie, la fase *Isolation* ($H = 4.168$) si distingue con un'entropia anche qui vicina al massimo teorico (4.585) e i tag *Compliment* ($H = 2.726$) e *Relationship* ($H = 2.766$) confermano quanto rilevato dall'analisi dell'entropia sulle emozioni primarie. Anche tra le fasi, *Fulfilling Needs* ($H = 2.726$) è la meno variabile, mentre *Isolation* ($H = 4.168$) conferma la sua complessità emotiva. Anche l'analisi degli stati affettivi, conferma i pattern identificati: i tag *Compliment* e *Relationship* mostrano la più bassa entropia, riflettendo una forte concentrazione su una singola emozione, come analizzato nella sezione 4.4. Al contrario, i tag *Communicative Desensitization* e *Approach* presentano entropia elevata, indicando una distribuzione più varia delle emozioni associate a questi tag.

Tra le fasi, *Fulfilling Needs* ha la più bassa entropia mentre *Isolation* si avvicina al massimo teorico di 3 bit, riflettendo una distribuzione quasi

uniforme.

Negli istogrammi nelle immagini 4.25 e 4.26, si propone una vista d'insieme dei risultati appena discussi, mostrando come l'entropia in *Fulfilling Needs* per le fasi e in *Compliment/Relationship* per i tag, sia minima; *Isolation*, invece, spicca in entrambi i casi, seguita da *Sexualizing Relationship* ma in misura minore.

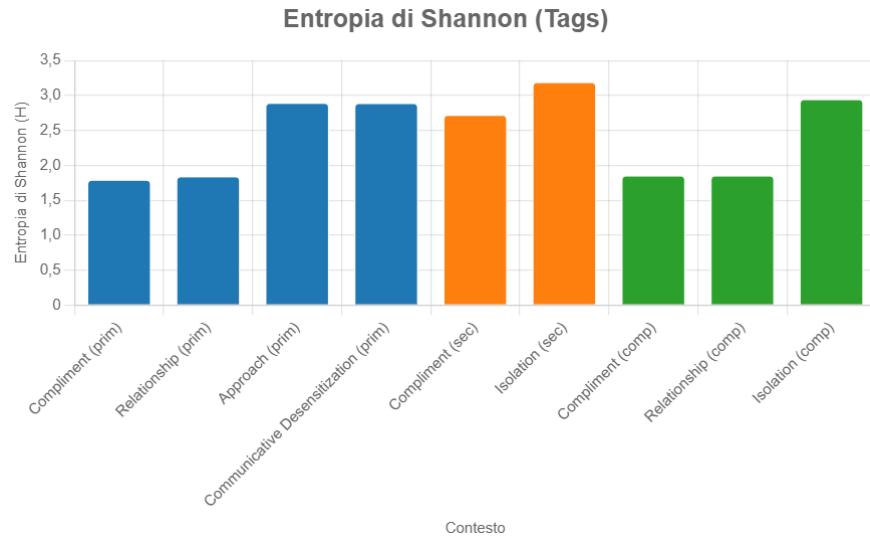


Figura 4.25: Entropia per Tags

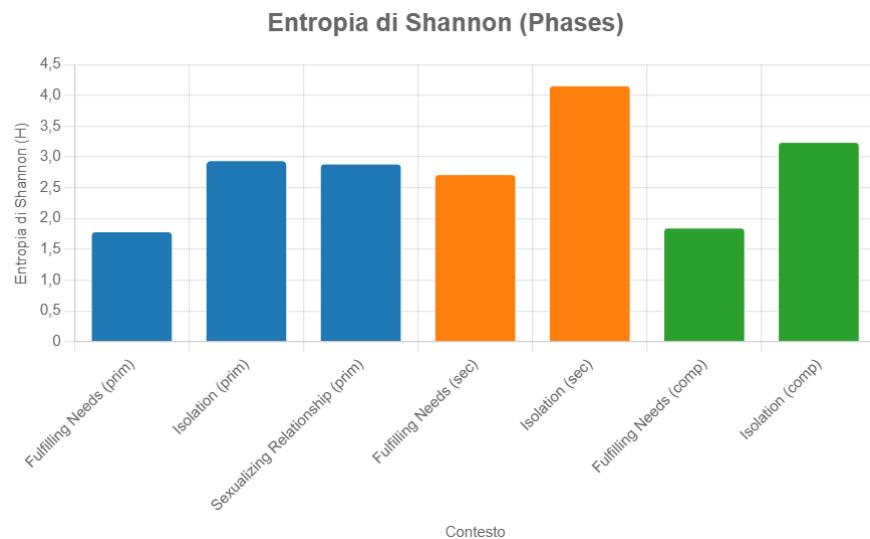


Figura 4.26: Entropia per Phases

Probabilità di Transizione

Questa analisi quantifica le probabilità di transizione (v. sez. 3.5) tra emozioni nel data set, considerando esclusivamente le fasi (*Targeting and Gaining Trust* → *Fulfilling Needs* → *Sexualizing Relationship* → *Isolation*), descritte nella sezione 1.1. In questo step di valutazione, non è stata effettuata l'analisi delle transizioni tra i tag emotivi, in quanto i tag non seguono un ordine stabile e possono comparire in più fasi diverse del processo. Questo rende difficile associare in modo chiaro un tag a una fase specifica. Di conseguenza, l'analisi delle transizioni tra tag avrebbe fornito risultati poco affidabili e con un alto grado di variabilità, non legati a nessuno schema dimostrabile. La tabella 4.27, riporta le probabilità di transizione maggiori di 0.3 (i.e., soglia minima) tra emozioni, associate alle rispettive fasi.

Tabella 4.27: Probabilità di Transizione per Emozioni Primarie (Fasi)

Transizione	Fasi	Probabilità (P)
trust → joy	Targeting and Gaining Trust → Fulfilling Needs	0.573
anticipation → joy	Targeting and Gaining Trust → Fulfilling Needs	0.554
surprise → joy	Targeting and Gaining Trust → Fulfilling Needs	0.551
sadness → joy	Targeting and Gaining Trust → Fulfilling Needs	0.547
surprise → sadness	Sexualizing Relationship → Isolation	0.471
fear → sadness	Sexualizing Relationship → Isolation	0.468
joy → surprise	Fulfilling Needs → Sexualizing Relationship	0.436
joy → fear	Fulfilling Needs → Sexualizing Relationship	0.434
trust → surprise	Fulfilling Needs → Sexualizing Relationship	0.342
trust → trust	Targeting and Gaining Trust → Fulfilling Needs	0.341
trust → fear	Fulfilling Needs → Sexualizing Relationship	0.340
anticipation → surprise	Fulfilling Needs → Sexualizing Relationship	0.339
anticipation → sadness	Sexualizing Relationship → Isolation	0.338
anticipation → fear	Fulfilling Needs → Sexualizing Relationship	0.337
anticipation → trust	Targeting and Gaining Trust → Fulfilling Needs	0.335
surprise → trust	Targeting and Gaining Trust → Fulfilling Needs	0.332
sadness → trust	Targeting and Gaining Trust → Fulfilling Needs	0.330

Si nota che da *Targeting and Gaining Trust* a *Fulfilling Needs*, la sequenza

emotiva principale è da *trust* a *joy* ($P=0.573$), mentre da *Sexualizing Relationship* a *Isolation* è da *surprise* a *sadness* ($P=0.471$). Tra le fasi *Fulfilling Needs* e *Sexualizing Relationship*, la sequenza principale rilevata è quella tra *joy* e *surprise* ($P=0.436$).

4.5 Conclusioni

La presente tesi ha esplorato l'applicazione dei modelli linguistici di grandi dimensioni (LLM) [33] nella prevenzione dell'adescamento online [10], con un focus particolare sullo sviluppo di un sistema per il riconoscimento delle emozioni [40]. In questa sezione conclusiva si propone una sintesi degli obiettivi raggiunti, del percorso metodologico adottato e dei principali contributi della ricerca, al fine di restituire una visione complessiva del lavoro svolto. Inoltre, vengono discussi i principali risultati ottenuti, analizzandone i limiti e i possibili sviluppi futuri in ambito scientifico e applicativo.

Sintesi dei risultati

Un contributo significativo è rappresentato dalla creazione del data set sintetico etichettato, descritto nella sezione 2.5 e basato sul modello emotionale di Plutchik, generato tramite tecniche di In-Context Learning [113] e Prompting [114], utilizzando i modelli linguistici Llama2-uncensored [145] e GPT [61]. Il data set, unico per la copertura dell'intero spettro emotivo di Plutchik, rappresenta una risorsa versatile e affidabile per task di emotion recognition [69], semantic similarity [44] e altre applicazioni di elaborazione del linguaggio naturale (NLP) [46]. L'approccio di transfer learning [106], applicato a diversi LLM e ottimizzato tramite fine-tuning [105] con iperparametri calibrati, ha evidenziato differenze prestazionali tra i vari modelli analizzati. In particolare, RoBERTa [122], si è distinto per accuratezza nella classificazione delle emozioni, confermando la validità dell'architettura e del processo di ottimizzazione adottato.

Nel task di similarità semantica, l’analisi comparativa tra modelli e metriche, ha evidenziato la superiorità del modello *Sentence-t5* [123] combinato con la metrica *FAISS Cosine Similarity* [85]. Il modello gerarchico completo, descritto nella sezione 3.3, ha mostrato buone prestazioni nella classificazione delle emozioni primarie e secondarie, evidenziando tuttavia prestazioni inferiori per gli stati affettivi composti, analizzati nella sezione 3.4.

Infine, l’applicazione del classificatore al data set sintetico di grooming [19], ha permesso l’identificazione di pattern emotivi salienti associati a ciasuna fase o tag (v. sez. 2.5). In particolare, la distribuzione delle emozioni e l’insieme delle probabilità di transizione, identificate nella sezione 4.4, costituiscono evidenze significative nel rilevamento precoce dei comportamenti predatori online.

Limiti e sviluppi futuri

Nonostante i risultati promettenti e la validità dell’approccio proposto, la ricerca presenta alcune limitazioni che ne circoscrivono l’ambito applicativo. Una prima criticità riguarda l’utilizzo esclusivo di dati sintetici generati in modo controllato per lo sviluppo e la validazione del sistema. Sebbene ciò assicuri coerenza e completezza del data set, tali dati potrebbero non riflettere pienamente la complessità, la varietà linguistica e le sfumature contestuali tipiche delle interazioni reali. Sviluppi futuri dovrebbero dunque prevedere l’applicazione e la verifica del sistema su dati reali, al fine di valutarne l’efficacia in contesti operativi concreti e aumentarne la robustezza.

Un’ulteriore limitazione riguarda la qualità delle etichette emotive del data set di addestramento (v. sez. 2.5), ottenute tramite tecniche di *In-Context Learning* e *Prompting* basate sul modello di Plutchik. Nonostante la validazione sia stata condotta su base quantitativa e qualitativa, la presenza di esempi semanticamente ambigui o emotivamente sovrapposti potrebbe aver compromesso la precisione del sistema di classificazione. Per ovviare a

tale problema nel processo di sviluppo potrebbe essere inclusa una fase di revisione manuale parziale per migliorare la qualità delle etichette.

L’analisi dei pattern emotivi ricorrenti, invece, condotta esclusivamente sul data set sintetico di casi di grooming, necessita di conferma su corpora reali per verificarne la validità e la generalizzabilità. Sviluppi futuri in questa direzione dovrebbero mirare alla costruzione o all’accesso a raccolte di dati autentici, eventualmente in collaborazione con enti istituzionali o piattaforme digitali, nel rispetto delle normative etiche e della privacy, verificando se i pattern rilevati utilizzando i data sets sintetici, sono corrispondenti a quelli ottenuti applicando il modello ai dati reali. Inoltre, il calcolo delle metriche, nonostante l’attenzione posta, potrebbe comprendere piccoli errori di approssimazione: si consiglia pertanto di verificare sempre il dato per un utilizzo più specifico.

Il modello gerarchico impiegato ha mostrato particolari difficoltà nella rilevazione degli stati affettivi, evidenziando una certa rigidità nel trattare emozioni sfumate o complesse. Per affrontare questa limitazione, sarà utile esplorare approcci alternativi, come architetture neurali più sofisticate, tecniche di multi-label classification o metriche valutative più adatte alla complessità del task.

Per incrementare la capacità del sistema di rilevare automaticamente casi di grooming in ambienti reali, sarà fondamentale integrare l’analisi emozionale con altre tecniche NLP complementari, quali il *Named Entity Recognition (NER)* [68]. Questa integrazione permetterebbe di identificare entità rilevanti come nomi propri, luoghi e riferimenti personali, arricchendo la comprensione del contesto e facilitando il riconoscimento di pattern manipolativi specifici, ad esempio la diffusione o la richiesta di informazioni personali.

Un ulteriore potenziamento potrebbe derivare dall’inclusione di analisi temporali, come la frequenza dei messaggi o le fasce orarie di invio, elementi spesso indicativi nei casi di adescamento, che possono offrire segnali

comportamentali utili al rilevamento automatico.

Dal punto di vista applicativo, invece, una prospettiva di particolare rilievo riguarda lo sviluppo di chatbot educativi per ambienti scolastici, con finalità didattiche e di sensibilizzazione. Questi strumenti, progettati per operare in modo anonimo e sicuro, potrebbero fornire ai minori un primo supporto nella comprensione e nel riconoscimento di situazioni potenzialmente pericolose, promuovendo al contempo la segnalazione tempestiva di comportamenti sospetti.

Infine, a partire dai risultati ottenuti in questo studio, una possibile evoluzione potrebbe consistere nel calcolo delle probabilità di transizione congiunte lungo intere sequenze di fasi e non solo tra coppie di fasi adiacenti. Ciò permetterebbe di individuare pattern più strutturati e ricorrenti, offrendo una rappresentazione più robusta e realistica delle dinamiche dell'adescamento.

Discussioni finali

Il lavoro presentato ha permesso di esplorare in modo concreto le potenzialità e i limiti dei *Large Language Models* (LLM) [33] in contesti critici, ponendo l'accento sulla necessità di bilanciare l'affidabilità dei risultati con la qualità e la disponibilità dei dati. In particolare, l'ambito della classificazione emozionale ha evidenziato la complessità intrinseca nel modellare stati affettivi composti, pur registrando buone prestazioni nel riconoscimento delle emozioni primarie e delle relative intensità.

L'assenza, nei principali repository pubblici, di modelli addestrati secondo il paradigma teorico di *Plutchik*, ha reso l'implementazione sviluppata un contributo rilevante, potenzialmente utile come base per ulteriori ricerche e miglioramenti. La fase sperimentale ha confermato l'efficacia del *transfer learning*, evidenziando al contempo l'importanza di una valutazione critica anche degli esiti negativi, spesso più indicativi dei limiti strutturali delle soluzioni adottate rispetto ai soli risultati soddisfacenti.

L’analisi della correlazione tra emozioni e fasi (o tag) dell’adescamento ha messo in evidenza, nella maggior parte dei casi, caratteristiche distintive coerenti e logicamente riconducibili ai comportamenti tipici descritti nei modelli teorici presenti in letteratura (v. sez. 1.1). Tali risultati, seppur promettenti, richiedono tuttavia una validazione più ampia e approfondita attraverso ulteriori sperimentazioni, come discusso nella sezione 4.5.

Dal punto di vista metodologico, il progetto ha privilegiato un approccio ampio e comparativo, volto a esplorare diverse soluzioni e configurazioni, piuttosto che perseguire esclusivamente l’ottimizzazione di un singolo modello. Questa impostazione è stata adottata con l’intento di offrire una risorsa utile non solo per il presente studio, ma anche per ricerche future in un ambito ancora poco esplorato e in continua evoluzione.

In prospettiva, è auspicabile che modelli linguistici sempre più sofisticati possano essere integrati in sistemi di prevenzione dell’adescamento online, contribuendo in modo concreto alla protezione delle persone più vulnerabili. In tal senso, questo lavoro si propone come un primo passo, consapevole dei propri limiti, ma orientato a stimolare lo sviluppo di approcci sempre più efficaci, in un campo in cui l’adozione di strumenti automatizzati potrebbe rappresentare un punto di svolta nel panorama della prevenzione.

Capitolo 5

Ringraziamenti

Desidero dedicare questo spazio a tutte le persone che, con il loro sostegno, hanno contribuito alla realizzazione di questa tesi di ricerca. In primo luogo, un ringraziamento speciale va ai miei relatori e co-relatori: la Prof.ssa Valentina Franzoni, il Dott. Emanuele Florindi, il Dott. Mattia Polticchia e il Prof. Alfredo Milani. La loro disponibilità, competenza e costante presenza sono stati per me un riferimento fondamentale, accompagnandomi con attenzione e professionalità in ogni fase del lavoro, dalla progettazione iniziale fino alla consegna finale. Un grazie va anche ai miei colleghi universitari, con cui ho condiviso momenti significativi lungo questo percorso, fatti di confronto, crescita e supporto reciproco. Ringrazio i miei amici, che hanno saputo comprendere e accettare le mie assenze, soprattutto nell'ultimo periodo, dovute alla gestione simultanea degli impegni di studio e lavoro, rimanendo sempre presenti. Un pensiero va alla mia famiglia, per il sostegno incondizionato e la fiducia che mi avete sempre trasmesso. Mi avete permesso di vivere questi anni con serenità, sentendomi sempre ascoltato, senza pressioni e con la possibilità di esprimere me stesso in ogni circostanza. Grazie: non è scontato, né comune. Un ringraziamento speciale va alla mia fidanzata. Hai vissuto con me ogni incertezza, ogni timore e ogni momento di difficoltà. Con la tua presenza, i tuoi incoraggiamenti e la tua capacità di guardare sempre avanti, mi hai dato la forza necessaria per non mollare.

Nelle notti di studio, nei momenti di ansia e nelle fasi più dure, sei stata la mia costante. Per tutto questo e per molto altro, ti sarò sempre grato. Infine, dedico questa tesi a me stesso, ai sacrifici compiuti e alla determinazione che mi ha portato fin qui. Ne è valsa la pena.

Bibliografia

- [1] *Report Polizia di stato 2024*. URL: <https://www.poliziadistato.it/statics/40/2024-report-def.-sppsc.pdf>.
- [2] OSINT Industries Team. *Social Media Intelligence (SOCMINT) in Modern Investigations*. URL: <https://www.osint.industries/post/social-media-intelligence-socmint-in-modern-investigations>.
- [3] *Analisi profilo adescatore Save the Children*. URL: https://www.savethechildren.es/sites/default/files/2023-11/OnlineGrooming_ESP.pdf.
- [4] *Thorn condivisione contenuti online*. URL: https://info.thorn.org/hubfs/Research/2022_Online_Grooming_Report.pdf.
- [5] Emily A. Greene-Colozzi et al. «Experiences and Perceptions of Online Sexual Solicitation and Grooming of Minors: A Retrospective Report». In: (2020). URL: <https://www.tandfonline.com/doi/abs/10.1080/10538712.2020.1801938>.
- [6] M. Kaliannan et al. *Parents' Awareness on Online Predators*. 2021. URL: https://www.researchgate.net/publication/356554356_Parents'_Awareness_on_Online_Predators_Cyber_Grooming_Deterrence.
- [7] *GDPR*. URL: <https://gdpr-info.eu/>.
- [8] *PCPD Official site (Hong Kong)*. URL: <https://www.pcpd.org.hk/>.
- [9] *Online grooming detection:A comprehensive survey of child exploitation in chat logs*. 2022. URL: <https://doi.org/10.1016/j.knosys.2022.110039>.

- [10] *Europol, Definizione di grooming Internet Organised Crime Threat Assessment (IOCTA) 2021.* URL: <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2021>.
- [11] *Gazzetta ufficiale, articolo 609 undecies.* URL: https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.versione=2&art.idGruppo=61&art.flagTipoArticolo=1&art.codiceRedazionale=030U1398&art.idArticolo=609&art.idSottoArticolo=11&art.idSottoArticolo1=10&art.dataPubblicazioneGazzetta=1930-10-26&art.progressivo=0.
- [12] *The 6 steps of online's grooming.* URL: <https://image.guardian.co.uk/sys-files/Society/documents/2003/07/17/Groomingreport.pdf>.
- [13] *Everyone deserves to be happy and safe(2017).* URL: https://www.researchgate.net/publication/326827010_Everyone_deserves_to_be_happy_and_safe_A_mixed_methods_study_exploring_how_online_and_offline_child_sexual_abuse_impact_young_people_and_how_professionals_respond_to_it.
- [14] *Offender strategies for engaging children in online sexual activity(2021).* URL: <https://www.sciencedirect.com/science/article/pii/S0145213421002878/pdfft?md5=ddedbad713c6944829aaa5a60ae14c81&pid=1-s2.0-S0145213421002878-main.pdf>.
- [15] *Offense Processes of Online Sexual Grooming and Abuse of Children Via Internet Communication Platforms.* 2017. URL: https://www.researchgate.net/publication/318361172_Offense_Processes_of_Online_Sexual_Grooming_and_Abuse_of_Children_Via_Internet_Communication_Platforms.
- [16] *CyberTipline 2022 Report.* URL: https://www.missingkids.org/test/c_y_b_e_r_t_i_p_l_i_n_e_d_a_t_a-draft.

- [17] Lillian Darke, Helen Paterson e Celine van Golde. «Illuminating Gaslighting: A Comprehensive Interdisciplinary Review of Gaslighting Literature». In: *Journal of Family Violence* (gen. 2025), pp. 1–17. DOI: 10.1007/s10896-025-00805-4.
- [18] *Il doppio legame*. URL: <https://www.torinopsicologo.com/il-doppio-legame/>.
- [19] Ludovico Guercio. «Generazione Sintetica chat di grooming». In: (2025).
- [20] Zhaoyang Niu, Guoqiang Zhong e Hui Yu. «A review on the attention mechanism of deep learning». In: *Neurocomputing* 452 (2021), pp. 48–62. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2021.03.091>. URL: <https://www.sciencedirect.com/science/article/pii/S092523122100477X>.
- [21] Pamela J. Black et al. «A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world». In: (2014). URL: <https://www.sciencedirect.com/science/article/abs/pii/S0145213414004360>.
- [22] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval*. 2008. URL: <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>.
- [23] Corinna Cortes e Vladimir Vapnik. «Support-vector networks». In: *Machine Learning* 20 (1995), pp. 273–297. URL: <https://link.springer.com/content/pdf/10.1007/BF00994018.pdf>.
- [24] Tomas Mikolov et al. «Efficient Estimation of Word Representations in Vector Space». In: (2013). URL: <https://arxiv.org/pdf/1301.3781>.
- [25] *Glove: Global Vectors for Word Representation*. URL: https://www.researchgate.net/publication/284576917_Glove_Global_Vectors_for_Word_Representation.

- [26] Alfredo Milani, Giulio Biondi e Valentina Franzoni. «User to Vector: Encoding User Behavior From Co-Occurrence of Observations». In: *IEEE Access* 12 (2024), pp. 156020–156037. DOI: 10.1109/ACCESS.2024.3485553.
- [27] Alfredo Milani e Valentina Franzoni. «PMING Distance: A Collaborative Semantic Proximity Measure». In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. 2012, pp. 155–162. URL: <https://dl.acm.org/doi/abs/10.1109/WI-IAT.2012.226>.
- [28] Sepp Hochreiter e Jurgen Schmidhuber. «Long Short-Term Memory». In: *Neural Computation* 9.8 (1997), pp. 1735–1780. URL: https://www.researchgate.net/publication/13853244_Long_Short-Term_Memory.
- [29] Yoon Kim. «Convolutional Neural Networks for Sentence Classification». In: (2014). arXiv: 1408.5882 [cs.CL]. URL: <https://arxiv.org/abs/1408.5882>.
- [30] Valentina Franzoni, Giulio Biondi e Alfredo Milani. «Morpho-Phraseological Based Classification of CEFR Italian L2 Learner Writing Proficiency». In: *IEEE Access* 12 (2024), pp. 156433–156441. DOI: 10.1109/ACCESS.2024.3485988.
- [31] Ashish Vaswani et al. «Attention Is All You Need». In: (2017). arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [32] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: (2019). arXiv: 1810.04805 [cs.CL]. URL: <https://arxiv.org/abs/1810.04805>.
- [33] Javier Ferrando et al. *A Primer on the Inner Workings of Transformer-based Language Models*. 2024. arXiv: 2405.00208 [cs.CL]. URL: <https://arxiv.org/abs/2405.00208>.
- [34] Sumit Chopra, Raia Hadsell e Yann Lecun. *Learning a similarity metric discriminatively, with application to face verification*. 2005.

- URL: https://www.researchgate.net/publication/4156225_Learning_a_similarity_metric_discriminatively_with_application_to_face_verification2.
- [35] *Clustering and Social Network Analysis*. 2020. URL: <https://www.cambridge.org/core/books/practical-data-science-for-information-professionals/clustering-and-social-network-analysis/1A997758F1534600356F9DAFA57B88E5>.
- [36] *Enhancing Privacy in the Early Detection of Sexual Predators Through Federated Learning and Differential Privacy*. 2025. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/35005/37160>.
- [37] Prosser e Edwards. *Helpful or Harmful? Exploring the Efficacy of Large Language Models for Online Grooming Prevention*. 2024. URL: <https://arxiv.org/abs/2403.09795#~:text=In%20this%20paper,%20the%20efficacy%20of%20LLMs%20for,by%20varying%20the%20provided%20context%20and%20prompt%20specificity..>
- [38] Ringenberg Bihani e Rayz. *Evaluating Language Models on Grooming Risk Estimation Using Fuzzy Theory*. 2025. URL: <https://arxiv.org/abs/2502.12563>.
- [39] PAUL EKMAN, *storia dei primi esperimenti sulle espressioni facciali con approccio scientifico contemporaneo*. URL: <https://labcns.com/articoli/corso-paul-ekman-espressioni-facciali-approccio-scientifico-tecniche-e-metodi-2/>.
- [40] Robert Plutchik e la teoria psicoevoluzionistica delle emozioni. 2017. URL: <https://it.manuelcappello.com/2017/08/robert-plutchik-e-la-teoria-psicoevoluzionistica-delle-emozioni/>.
- [41] Ibrahim Al Azher et al. *FutureGen: LLM-RAG Approach to Generate the Future Work of Scientific Article*. 2025. arXiv: 2503.16561 [cs.CL]. URL: <https://arxiv.org/abs/2503.16561>.

- [42] Karl Moritz Hermann. *Distributed Representations for Compositional Semantics*. 2014. arXiv: 1411.3146 [cs.CL]. URL: <https://arxiv.org/abs/1411.3146>.
- [43] H. Meijer, J. Truong e R. Karimi. *Document Embedding for Scientific Articles: Efficacy of Word Embeddings vs TFIDF*. Lug. 2021. doi: 10.48550/arXiv.2107.05151.
- [44] Goutam Majumder et al. «Semantic Textual Similarity Methods, Tools, and Applications: A Survey». In: *Computacion y Sistemas* 20 (dic. 2016), pp. 647–665. doi: 10.13053/CyS-20-4-2506.
- [45] Marco Viola. «Rappresentazioni scientifiche dell'emotività: dalle emozioni di base al core affect (... e oltre?)» In: set. 2021, pp. 151–176. ISBN: 978-88-7885-994-4.
- [46] Supriyono et al. «Advancements in natural language processing: Implications, challenges, and future directions». In: *Telematics and Informatics Reports* 16 (2024), p. 100173. ISSN: 2772-5030. doi: <https://doi.org/10.1016/j.teler.2024.100173>. URL: <https://www.sciencedirect.com/science/article/pii/S2772503024000598>.
- [47] *LINGUISTICA COMPUTAZIONALE*. URL: <https://www.ilc.cnr.it/lingistica-computazionale/>.
- [48] *Intelligenza artificiale*. URL: https://it.wikipedia.org/wiki/Intelligenza_artificiale.
- [49] *Machine Learning*. URL: https://en.wikipedia.org/wiki/Machine_learning.
- [50] *DIRITTO PENALE* *Violenza psicologica e manipolazione integrano reato?* URL: <https://www.studiolegalearenosto.it/diritto-penale-violenza-psicologica-e-manipolazione-integrano-reato/>.
- [51] *Natural Language Processing*. URL: <https://huggingface.co/learn/llm-course/it/chapter1/2>.

- [52] MR RAJESH e Tryambak Hiwarkar. «Exploring Preprocessing Techniques for Natural LanguageText: A Comprehensive Study Using Python Code». In: *international journal of engineering technology and management sciences* 7 (gen. 2023), pp. 390–399. DOI: 10.46647/ijetms.2023.v07i05.047.
- [53] Qi Liu, Matt J. Kusner e Phil Blunsom. *A Survey on Contextual Embeddings*. 2020. arXiv: 2003 . 07278 [cs.CL]. URL: <https://arxiv.org/abs/2003.07278>.
- [54] Yang Zhang et al. *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. 2025. arXiv: 2403 . 02901 [cs.AI]. URL: <https://arxiv.org/abs/2403.02901>.
- [55] *Understanding Tokenization, Stemming, and Lemmatization in NLP*. URL: <https://becominghuman.ai/understanding-tokenization-stemming-and-lemmatization-in-nlp-ba7944bb92a0>.
- [56] *Sparse and Dense Embeddings*. URL: <https://zilliz.com/learn/sparse-and-dense-embeddings>.
- [57] *Word2Vec – Rappresentazione Vettoriale Semantica delle Parole Python 8 min lettura*. URL: <https://zilliz.com/learn/sparse-and-dense-embeddings>.
- [58] *GloVe: Global Vectors for Word Representation*. URL: <https://nlp.stanford.edu/projects/glove/>.
- [59] *FastText, Library for efficient text classification and representation learning*. URL: <https://fasttext.cc/>.
- [60] *bert-base-uncased*. URL: <https://huggingface.co/google-bert/bert-base-uncased>.
- [61] OpenAI. *ChatGPT*. URL: <https://chatgpt.com/>.
- [62] *CBOW (Continuous Bag of words)*. URL: <https://towardsmachinelearning.org/cbow-continuous-bag-of-words/>.

- [63] *Cosa significa "Skip-Gram"?* URL: <https://scisimple.com/it/keywords/skip-gram--kkxd6g0>.
- [64] *ELMo*. URL: <https://en.wikipedia.org/wiki/ELMo>.
- [65] Alladi Deekshith. «ADVANCES IN NATURAL LANGUAGE PROCESSING: A SURVEY OF TECHNIQUES». In: *International Journal of Innovations in Engineering Research and Technology* 8 (ott. 2024), pp. 74–83. DOI: 10.26662/ijiert.v8i3.pp74-83.
- [66] Dibyanshu Bhatt. *Self-Attention to State-of-the-Art: An Exploration of Transformer Architecture and Applications*. Dic. 2024. DOI: 10.13140/RG.2.2.33498.04807.
- [67] Shreya Acharya et al. «Question Answering System using NLP and BERT». In: *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. 2022, pp. 925–929.
- [68] *Cos'è la named entity recognition?* URL: <https://www.ibm.com/it-it/think/topics/named-entity-recognition>.
- [69] Nikolaos Mylonas et al. «Online Child Grooming Detection: Challenges and Future Directions». In: set. 2024, pp. 237–247. ISBN: 978-3-031-62082-9. DOI: 10.1007/978-3-031-62083-6_19.
- [70] Rosalind W. Picard. *Affective Computing*. Cambridge, MA: MIT Press, 1997.
- [71] Zein Samira e Yodit Weldegeorgise. «A Systematic Review of Deep Learning Methods: Classification, Selection, and Scientific Understanding». In: *International Journal of Sciences Basic and Applied Research (IJSBAR)* 74 (dic. 2024), pp. 145–153.
- [72] Robert Plutchik. «A General Psychoevolutionary Theory of Emotion». In: *Theories of Emotion* 1 (1980), pp. 3–33.
- [73] David Ortiz-Perez et al. «Multimodal Fusion Strategies for Emotion Recognition». In: *2024 International Joint Conference on Neural Networks (IJCNN)*. 2024, pp. 1–8.

- [74] Paul Ekman. «An Argument for Basic Emotions». In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10.1080/02699939208411068.
- [75] Jiawei Han, Micheline Kamber e Jian Pei. «Getting to Know Your Data». In: 2012. URL: <https://www.com/science/article/pii/B9780123814791000022>.
- [76] *Euclidean Distance*. URL: <https://www.sciencedirect.com/topics/mathematics/euclidean-distance#definition>.
- [77] *Jaccard*. URL: <https://www.statology.org/jaccard-similarity/>.
- [78] *Sørensen-Dice Coefficient: A Comprehensive Guide to Similarity Measurement*. URL: <https://researchdatapod.com/sorensendice-coefficient/>.
- [79] *Jaccard Similarity and k-Grams*. URL: <https://users.cs.utah.edu/~jeffp/DMBook/L3-Jaccard+nGram.pdf>.
- [80] *sokalsneath*. URL: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.sokalsneath.html>.
- [81] *sokalSneath2: Sokal Sneath Index 2*. URL: <https://rdrr.io/cran/partitionComparison/man/sokalSneath2.html>.
- [82] *Rogers Tanimoto Distance*. URL: <https://distancia.readthedocs.io/en/latest/RogersTanimoto.html>.
- [83] Pang-Ning Tan, Michael Steinbach e Vipin Kumar. *Introduction to Data Mining*. Boston: Pearson Addison Wesley, 2005.
- [84] Mengzhao Wang et al. *A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search*. 2021. arXiv: 2101.12631 [cs.IR]. URL: <https://arxiv.org/abs/2101.12631>.
- [85] *Faiss*. URL: <https://python.langchain.com/docs/integrations/vectorstores/faiss/>.
- [86] Giulio Ermanno Pibiri e Rossano Venturini. «Techniques for Inverted Index Compression». In: *ACM Computing Surveys* 53.6 (dic. 2020),

- pp. 1–36. ISSN: 1557-7341. DOI: 10.1145/3415148. URL: <http://dx.doi.org/10.1145/3415148>.
- [87] *Struct faiss::IndexPQ*. URL: https://faiss.ai/cpp_api/struct/structfaiss_1_1IndexPQ.html.
- [88] *Struct faiss::IndexFlatL2*. URL: https://faiss.ai/cpp_api/struct/structfaiss_1_1IndexFlatL2.html.
- [89] *SentenceTransformers Documentation*. URL: <https://www.sbert.net/>.
- [90] *What is the k-nearest neighbors (KNN) algorithm?* URL: <https://www.ibm.com/think/topics/knn>.
- [91] Alladi Deekshith. «ADVANCES IN NATURAL LANGUAGE PROCESSING: A SURVEY OF TECHNIQUES». In: *International Journal of Innovations in Engineering Research and Technology* 8 (ott. 2024), pp. 74–83. DOI: 10.26662/ijiert.v8i3.pp74-83.
- [92] Or Peretz, Michal Koren e Oded Koren. «Naive Bayes classifier – An ensemble procedure for recall and precision enrichment». In: *Engineering Applications of Artificial Intelligence* 136 (2024), p. 108972. ISSN: 0952-1976. DOI: <https://doi.org/10.1016/j.engappai.2024.108972>. URL: <https://www.sciencedirect.com/science/article/pii/S0952197624011308>.
- [93] *Teorema di bayes*. URL: https://www.bayes.it/bayesarc/html/teorema_di_bayes.html.
- [94] *Properties of Naive Bayes*. 2008. URL: <https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html>.
- [95] Jair Cervantes et al. «A comprehensive survey on support vector machine classification: Applications, challenges and trends». In: *Neurocomputing* 408 (2020), pp. 189–215. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2019.10.118>. URL: <https://www.sciencedirect.com/science/article/pii/S0925231220307153>.

- [96] *Efficiency of SVM classifier with Word2Vec and Doc2Vec models*. URL: <https://intapi.sciendo.com/pdf/10.2478/icas-2019-0043>.
- [97] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. 2015. arXiv: 1407.7502 [stat.ML]. URL: <https://arxiv.org/abs/1407.7502>.
- [98] *Varianza*. URL: <https://www.analisi-statistica.com/varianza/>.
- [99] *Cosa sono l'Overfitting e l'Underfitting?* URL: <https://www.intelligenzaartificialei.net/post/cosa-sono-l-overfitting-e-l-underfitting-e-come-puoi-limitarli-nel-machine-learning>.
- [100] Sedir Mohammed et al. «The effects of data quality on machine learning performance on tabular data». In: *Information Systems* 132 (lug. 2025), p. 102549. ISSN: 0306-4379. DOI: 10.1016/j.is.2025.102549. URL: <http://dx.doi.org/10.1016/j.is.2025.102549>.
- [101] Prasenjit Dey, Srujana Merugu e Sivaramakrishnan Kaveri. «Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models». In: *Companion Proceedings of the ACM on Web Conference 2025*. WWW '25. ACM, mag. 2025, pp. 947–951. DOI: 10.1145/3701716.3715523. URL: <http://dx.doi.org/10.1145/3701716.3715523>.
- [102] Domor Mienye, Theo Swart e George Obaido. «Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications». In: *Information* 15 (ago. 2024), p. 517. DOI: 10.3390/info15090517.
- [103] Alexander Rehmer e Andreas Kroll. «On the vanishing and exploding gradient problem in Gated Recurrent Units». In: *IFAC-PapersOnLine* 53.2 (2020). 21st IFAC World Congress, pp. 1243–1248. ISSN: 2405-8963. DOI: <https://doi.org/10.1016/j.ifacol.2020.12.1342>. URL: <https://www.sciencedirect.com/science/article/pii/S2405896320317481>.

- [104] *Gating mechanism*. URL: https://en.wikipedia.org/wiki/Gating_mechanism.
- [105] Samar Pratap et al. «The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques». In: *Natural Language Processing Journal* 11 (2025), p. 100144. ISSN: 2949-7191. DOI: <https://doi.org/10.1016/j.nlp.2025.100144>. URL: <https://www.sciencedirect.com/science/article/pii/S2949719125000202>.
- [106] Asmaul Hosna et al. «Transfer learning: a friendly introduction». In: *Journal of Big Data* 9 (ott. 2022). DOI: [10.1186/s40537-022-00652-w](https://doi.org/10.1186/s40537-022-00652-w).
- [107] Juan Terven et al. «A comprehensive survey of loss functions and metrics in deep learning». In: *Artificial Intelligence Review* 58.7 (2025). ISSN: 1573-7462. DOI: [10.1007/s10462-025-11198-7](https://doi.org/10.1007/s10462-025-11198-7). URL: <http://dx.doi.org/10.1007/s10462-025-11198-7>.
- [108] Everton L. Aleixo et al. *Catastrophic Forgetting in Deep Learning: A Comprehensive Taxonomy*. 2023. arXiv: 2312.10549 [cs.LG]. URL: <https://arxiv.org/abs/2312.10549>.
- [109] Jian Gu et al. *A Semantic-Aware Layer-Freezing Approach to Computation-Efficient Fine-Tuning of Language Models*. 2025. arXiv: 2406.11753 [cs.CL]. URL: <https://arxiv.org/abs/2406.11753>.
- [110] Krishna Prasad Varadarajan Srinivasan et al. *Comparative Analysis of Different Efficient Fine Tuning Methods of Large Language Models (LLMs) in Low-Resource Setting*. 2024. arXiv: 2405.13181 [cs.CL]. URL: <https://arxiv.org/abs/2405.13181>.
- [111] Kshitij Gupta et al. *Continual Pre-Training of Large Language Models: How to (re)warm your model?* 2023. arXiv: 2308.04014 [cs.CL]. URL: <https://arxiv.org/abs/2308.04014>.
- [112] Jesse Dodge et al. *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*. 2020. arXiv: 2002.06305 [cs.CL]. URL: <https://arxiv.org/abs/2002.06305>.

- [113] Dong Qingxiu et al. «A Survey on In-context Learning». In: (2022). arXiv: 1408.5882 [cs.CL]. URL: <https://arxiv.org/abs/2301.00234>.
- [114] Shuofei Qiao et al. *Reasoning with Language Model Prompting: A Survey*. 2023. arXiv: 2212.09597 [cs.CL]. URL: <https://arxiv.org/abs/2212.09597>.
- [115] *F1 score*. URL: https://it.wikipedia.org/wiki/F1_score.
- [116] *Precisione*. URL: <https://it.wikipedia.org/wiki/Precisione>.
- [117] *Richiamo*. URL: <https://it.wikipedia.org/wiki/Richiamo>.
- [118] *Curva ROC*. URL: https://es.wikipedia.org/wiki/Curva_ROC.
- [119] *Cos'è: Area sotto curva (AUC)*. URL: <https://it.statisticseasilly.com/glossario/cos%27%C3%A8-1%27area-sotto-la-curva-auc/>.
- [120] *Confusion matrix*. URL: https://en.wikipedia.org/wiki/Confusion_matrix.
- [121] *Accuracy*. URL: <https://it.wikipedia.org/wiki/Accuratezza>.
- [122] *RoBERTa*. URL: https://huggingface.co/docs/transformers/model_doc/roberta.
- [123] Jianmo Ni et al. *Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models*. 2021. arXiv: 2108.08877 [cs.CL]. URL: <https://arxiv.org/abs/2108.08877>.
- [124] *Understanding Encoder And Decoder LLMs*. URL: <https://sebastianraschka.com/p/understanding-encoder-and-decoder>.
- [125] *Modelli encoder*. URL: <https://huggingface.co/learn/llm-course/it/chapter1/5>.
- [126] *Modelli Decoder*. URL: <https://huggingface.co/learn/llm-course/it/chapter1/6>.
- [127] Crescenzio Gallo e Michelangelo De Bonis. *Reti Neurali Artificiali - Tutorial*. Gen. 2016.

- [128] Bing Xu et al. *Empirical Evaluation of Rectified Activations in Convolutional Network*. 2015. arXiv: 1505.00853 [cs.LG]. URL: <https://arxiv.org/abs/1505.00853>.
- [129] Zhihao Fan et al. *Mask Attention Networks: Rethinking and Strengthen Transformer*. 2021. arXiv: 2103.13597 [cs.CL]. URL: <https://arxiv.org/abs/2103.13597>.
- [130] Mozhdeh Gheini, Xiang Ren e Jonathan May. *Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation*. 2021. arXiv: 2104.08771 [cs.CL]. URL: <https://arxiv.org/abs/2104.08771>.
- [131] Rajendran Nirthika et al. «Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study». In: 34.7 (2022). ISSN: 0941-0643. URL: <https://doi.org/10.1007/s00521-022-06953-8>.
- [132] Hyunjin Choi et al. *Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks*. 2021. arXiv: 2101.10642 [cs.CL]. URL: <https://arxiv.org/abs/2101.10642>.
- [133] Kailash Hambarde e Hugo Proen  a. «Information Retrieval: Recent Advances and Beyond». In: *IEEE Access* PP (gen. 2023), pp. 1–1. DOI: [10.1109/ACCESS.2023.3295776](https://doi.org/10.1109/ACCESS.2023.3295776).
- [134] Jan Philip Wahle et al. «Identifying Machine-Paraphrased Plagiarism». In: *Information for a Better World: Shaping the Global Future*. Springer International Publishing, 2022, pp. 393–413. ISBN: 9783030969578. DOI: [10.1007/978-3-030-96957-8_34](https://doi.org/10.1007/978-3-030-96957-8_34). URL: http://dx.doi.org/10.1007/978-3-030-96957-8_34.
- [135] Byung-Rae Cha e Binod Vaidya. «Enhancing Human Activity Recognition with Siamese Networks: A Comparative Study of Contrastive and Triplet Learning Approaches». In: *Electronics* 13.9 (2024). URL: <https://www.mdpi.com/2079-9292/13/9/1739>.

- [136] Elad Hoffer e Nir Ailon. *Deep metric learning using Triplet network*. 2018. arXiv: 1412.6622 [cs.LG]. URL: <https://arxiv.org/abs/1412.6622>.
- [137] Luciano Serafini e Artur d'Avila Garcez. *Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge*. 2016. arXiv: 1606.04422 [cs.AI]. URL: <https://arxiv.org/abs/1606.04422>.
- [138] Mohammad Munzir Ahanger, Mohd Arif Wani e Vasile Palade. «sBERT: Parameter-Efficient Transformer-Based Deep Learning Model for Scientific Literature Classification». In: *Knowledge* 4.3 (2024), pp. 397–421. URL: <https://www.mdpi.com/2673-9585/4/3/22>.
- [139] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: <https://arxiv.org/abs/1910.01108>.
- [140] *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. URL: <https://huggingface.co/papers/2002.10957s>.
- [141] Alexis Conneau et al. *Unsupervised Cross-lingual Representation Learning at Scale*. 2020. arXiv: 1911.02116 [cs.CL]. URL: <https://arxiv.org/abs/1911.02116>.
- [142] *Albert*. URL: https://huggingface.co/docs/transformers/model_doc/albert.
- [143] *MPNet*. URL: https://huggingface.co/docs/transformers/model_doc/mpnet.
- [144] *Byte-Pair Encoding tokenization*. URL: <https://huggingface.co/learn/llm-course/chapter6/5>.
- [145] Meta. *llama2-uncensored*. URL: <https://ollama.com/library/llama2-uncensored>.
- [146] Luigi Villani e Psicologia. «I Bias Cognitivi». In: (set. 2023).

- [147] Ludovico Guercio. «Verso la Prevenzione del Grooming Online: Classificazione dell’Arousal nelle Prime Fasi dell’Adescamento Online». In: (2025).
- [148] *pandas*. URL: <https://pandas.pydata.org/>.
- [149] Rodionova Oxana, Kucheryavskiy Sergey e Pomerantsev Alexey. «Efficient tools for principal component analysis of complex data». In: (2021). URL: <https://www.sciencedirect.com/science/article/pii/S0169743921000721>.
- [150] *Addestramento con Dataset Sbilanciati*. URL: <https://arxiv.org/pdf/2008.09209.pdf>.
- [151] Roweida Mohammed, Jumanah Rawashdeh e Malak Abdullah. «Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results». In: apr. 2020, pp. 243–248. DOI: 10.1109/ICICS49469.2020.239556.
- [152] *llama3.1-8b-abliterated*. URL: <https://ollama.com/mannix/llama3.1-8b-abliterated>.
- [153] *Perverted Justice dataset, PAN12 Deception Detection: Sexual Predator Identification*. 2012. URL: <https://github.com/helenkoutli/PervertedJusticeDataset>.
- [154] Richard Meyes et al. *Ablation Studies in Artificial Neural Networks*. 2019. arXiv: 1901.08644 [cs.NE]. URL: <https://arxiv.org/abs/1901.08644>.
- [155] *LabelEncoder*. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>.
- [156] *Learning rate*. URL: https://en.wikipedia.org/wiki/Learning_rate.
- [157] *What is batch size in neural network?* URL: <https://stats.stackexchange.com/questions/153531/what-is-batch-size-in-neural-network>.

- [158] *Weight Decay*. URL: https://d2l.ai/chapter_linear-regression/weight-decay.html.
- [159] *Epoch (computing)*. URL: [https://en.wikipedia.org/wiki/Epoch_\(computing\)](https://en.wikipedia.org/wiki/Epoch_(computing)).
- [160] *Transformers*. URL: <https://huggingface.co/docs/transformers/index>.
- [161] *PyTorch*. URL: <https://pytorch.org/>.
- [162] *datasets*. URL: <https://pypi.org/project/datasets/>.
- [163] *Scikit-Learn*. URL: <https://scikit-learn.org/stable/index.html>.
- [164] *Numpy*. URL: <https://numpy.org/>.
- [165] *matplotlib*. URL: <https://matplotlib.org/>.
- [166] *seaborn: statistical data visualization*. URL: <https://seaborn.pydata.org/>.
- [167] *False positives and false negatives*. URL: https://en.wikipedia.org/wiki/False_positives_and_false_negatives.
- [168] *Analisi del sentiment*. URL: https://it.wikipedia.org/wiki/Analisi_del_sentiment.
- [169] *Preparing Text Data for Transformers: Tokenization, Mapping and Padding*. URL: <https://medium.com/@lokaregns/preparing-text-data-for-transformers-tokenization-mapping-and-padding-9fbfbce28028>.
- [170] *Trainer*. URL: https://huggingface.co/docs/transformers/main_classes/trainer.
- [171] Ilya Loshchilov e Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: <https://arxiv.org/abs/1711.05101>.
- [172] *Frequenza (statistica)*. URL: [https://it.wikipedia.org/wiki/Frequenza_\(statistica\)](https://it.wikipedia.org/wiki/Frequenza_(statistica)).

- [173] Wolfgang H. Berger e Frances L. Parker. «Diversity of Planktonic Foraminifera in Deep-Sea Sediments». In: *Science* 168.3937 (1970), pp. 1345–1347. DOI: 10.1126/science.168.3937.1345. URL: <https://www.science.org/doi/abs/10.1126/science.168.3937.1345>.
- [174] *Entropia (teoria dell'informazione)*. URL: [https://it.wikipedia.org/wiki/Entropia_\(teoria_dell%27informazione\)](https://it.wikipedia.org/wiki/Entropia_(teoria_dell%27informazione)).
- [175] *Processo markoviano*. URL: https://it.wikipedia.org/wiki/Processo_markoviano.
- [176] *Probabilità condizionata*. URL: https://it.wikipedia.org/wiki/Probabilit%C3%A0_condizionata.