



UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Laurea Triennale in Ingegneria Informatica

**Analisi comparativa di metodi di
autovalutazione delle emozioni tramite
tecniche di Data Mining**

Relatore:

Dr. Francesco Pistolesi

Ing. Michele Baldassini

Candidato:

Jonathan Frattacci

ANNO ACCADEMICO 2023/2024

Indice

1	Introduzione	7
1.1	Data Mining	7
1.1.1	Come funziona il data mining	7
1.1.2	Tecniche di data mining	8
1.1.3	Mining di regole Associative	9
2	Interpretazione delle emozioni	13
2.1	James Russell e la Classificazione delle emozioni	13
2.1.1	Modello PAD	14
2.1.2	Modello Circomplesso dell’Affetto	14
2.2	Limitazioni modelli ad approccio dimensionale	17
3	Affective Tagging e Data Collection	19
3.1	Affective Tagging	19
3.1.1	Il test	19
3.1.2	Il confronto	20
3.2	Il mio contributo	20
4	Analisi dei Dati e Conclusioni	21
4.1	Struttura dataset	21
4.2	Analisi Dataset	22
4.2.1	Suddivisione utenti in fasce d’età	22
4.2.2	Decisione valori adeguati	23
4.2.3	Applicazione algoritmo Apriori	25
4.2.4	Regole associative	27
4.2.5	Risultati	30
4.3	Conclusioni	32
4.3.1	Use Cases	33
	Ringraziamenti	35

Elenco delle figure

2.1	Scala di valutazione PAD: S.A.M.	15
2.2	Collocazione degli otto parametri sul piano Valence(Pleasure)-Arousal . .	16
2.3	Collocazione delle emozioni sul piano Valence-Arousal	16
2.4	Affective Slider	16
3.1	Tutorial mostrato ad inizio test	20
3.2	Piano di Russell rivisitato in ottica emoji	20
4.1	Struttura dataset	22
4.2	Istogramma intervalli da 8 anni	22
4.3	Suddivisione piano di Russell	23

Capitolo 1

Introduzione

Vivendo oggi in quella che viene definita da alcuni sociologi come "società dell'informazione", cioè una società nella quale beni immateriali come i dati hanno un valore di gran lunga superiore a quello dei beni materiali, diventa essenziale l'applicazione di metodi di Data Mining al fine di analizzare una grande quantità di informazioni.

1.1 Data Mining

Il termine *Data Mining* può portare ad avere una concezione errata di cosa questa branca dell'analisi dei dati permetta effettivamente di fare con essi. Dall'inglese "data" (dati) e "to mine" (estrarre), la letterale traduzione porta ad interpretare questo insieme di tecniche come «estrazione *di* dati», mentre una più corretta accezione sarebbe quella di «estrazione *dai* dati». Infatti, sebbene la raccolta dei dati sia una fase fondamentale, l'obiettivo principale di queste tecniche rimane quello di scoprire relazioni tra essi, tendenze e modelli, e ricavare informazioni utili.

1.1.1 Come funziona il data mining

Data la vastità degli ambiti in cui queste tecniche possono essere applicate, è lecito affermare che non è possibile definire una vera e propria procedura standard, sono al più definibili delle fasi organizzative e preparatorie che possano fungere da linee guida per i data miners[4]. In particolare esse sono:

1. Comprensione del problema o dell'area di indagine
2. Raccolta e Comprensione dei dati
3. Preparazione dei dati
 - (a) Pulizia dei dati
 - (b) Integrazione dei dati
 - (c) Formattazione dei dati
4. Modellazione dei dati

5. Valutazione dei dati
6. Implementazione e formazione degli utenti

1.1.2 Tecniche di data mining

Gli approcci di data mining possono essere suddivisi in due principali categorie:

1. **Approcci diretti.**

Focalizzati su un obiettivo specifico e ben definito. In questo caso, l'analisi è progettata per ottenere un risultato preciso e predefinito.

2. **Approcci indiretti.**

Concepiti come un processo di scoperta, mirano a esplorare i dati e identificare pattern inaspettati o relazioni nascoste. L'obiettivo non è predefinito rigidamente, ma si evolve durante l'analisi.

1.1.2.1 Tecniche ad approccio indiretto

In questa sezione verranno leggermente approfondite le tecniche ad approccio indiretto in quanto, nel lavoro spiegato nei capitoli successivi viene applicato solamente questo approccio.

Alcuni esempi sono:

- Rilevamento anomalie: identificare dati che si discostano significativamente dalla norma all'interno di un insieme di dati con modelli riconoscibili. Utile per individuare frodi, errori o eventi eccezionali.
- Regressione: ricerca di pattern che permettano di ottenere valori futuri dei dati a partire da insiemi di eventi o valori passati. Utile nel caso si vogliano riconoscere variazioni nel tempo dei dati raccolti.
- Clustering
- Classificazione dati
- Mining di regole di associazione

Clustering e Classificazione dati

La Classificazione è una tecnica di data mining che ha lo scopo di addestrare un algoritmo di Machine Learning affinché esso sia in grado, dato un insieme di categorie note, di riconoscere la categoria di appartenenza di un determinato dato e di assegnarglielo. Un esempio della sua applicazione potrebbe essere quello di un software che è stato addestrato a distinguere malattie sulla base dei sintomi accusati oppure che sappia distinguere l'immagine di un ortaggio rispetto ad un altro.

Il Clustering può sembrare molto simile concettualmente a quanto appena descritto, ma in realtà vi è una differenza che all'atto pratico è tutt'altro che trascurabile. Infatti, sebbene anche questa tecnica abbia lo scopo di raggruppare i dati in categorie, la differenza sta nel fatto che qua non esistono categorie note a priori né alcun tipo di etichetta.

Proprio per questo, mentre nella classificazione è necessario aver prima interpretato il dato per poi poterlo assegnare ad una categoria, in questo caso è possibile raggruppare i dati in *cluster* tramite caratteristiche comuni e interpretarli in un secondo momento. Un esempio di applicazione di questa seconda tecnica è quello in cui si vogliano raggruppare i clienti di un supermercato in categorie basate sui loro gusti di acquisto.

1.1.3 Mining di regole Associative

Si tratta di una tecnica di Data mining che mira a scoprire relazioni tra elementi apparentemente non correlati all'interno del dataset. Per ottenere tale risultato si sfrutta un concetto fondamentale di questa tecnica, quello di **insieme frequente** [2]. La definizione di insieme frequente è quella di "gruppo di elementi che compaiono frequentemente insieme in un dataset", ma come è possibile la frequenza di un insieme di elementi?

A questo scopo è definito il **supporto**, ovvero il numero di transazioni nel dataset all'interno delle quali compare l'insieme di elementi di cui si voglia misurare la frequenza. Esso rappresenta un importante parametro di valutazione, in quanto una determinata percentuale di supporto per una regola, rappresenterà allo stesso tempo la percentuale di transazioni all'interno del dataset che rispettano quella regola. Sebbene il supporto sia uno strumento necessario soprattutto nella fase di ricerca degli insiemi frequenti, esso non risulta essere sufficiente a scovare regole di associazione.

Viene quindi introdotto il concetto di **confidenza**, il quale, data una regola $A \rightarrow B$ (con A elemento o insieme di elementi e B elemento che appare spesso insieme ad A), è definito come:

$$Confidenza(A \rightarrow B) = \frac{supporto(A \cup B)}{supporto(A)}$$

Il terzo concetto importante da introdurre è quello di **lift**. Esso rappresenta il legame che esiste tra probabilità che un evento si verifichi e quanto ci si aspetta che ciò accada in modo casuale. Il valore limite è 1, valori di lift maggiori di 1 indicano che il verificarsi di un dato evento è più probabile rispetto al caso, mentre per valori inferiori significa che è meno probabile e viene calcolato come:

$$Lift(A \rightarrow B) = \frac{confidenza(A \rightarrow B)}{supporto(B)}$$

Tra i vantaggi di questa tecnica troviamo sicuramente:

- **facilità di interpretazione:** rispetto ad altre tecniche di mining, in questo caso risulta più semplice interpretare i pattern risultanti dall'analisi dei dati.
- **ampia applicabilità:** è possibile applicare gli algoritmi di mining di regole associative a svariati ambiti anche sconnessi tra loro (sicuramente tra i più intuitivi vi sono le ricerche di marketing da parte di business come supermercati, ma non solo).

- **grandi capacità di gestione:** questa tecnica è in grado di individuare con successo pattern anche in dataset di grandi dimensioni, rendendosi quindi adatta all'applicazione sui Big Data.

Valutando gli svantaggi si scopre che vi sono:

- **Limiti nella scoperta di relazioni complesse:** questa tecnica presenta svariati limitazioni come:
 1. Co-occorrenza: concentrandosi solo sulla co-occorrenza degli elementi all'interno di una stessa transazione, l'algoritmo è in grado di identificare relazioni solo tra gli elementi che compaiono insieme nella stessa.
 2. Mancanza di informazioni di sequenza: non tenendo conto dell'ordine di comparsa degli elementi in una transazione, non è possibile identificare relazioni dipendenti dal tempo o dalla causalità.
 3. Relazioni di dipendenza: l'algoritmo si limita ad identificare associazioni tra gli elementi ma non è in grado di distinguere casi in cui uno sia causato dall'altro.
 4. Mancanza di generalizzazione: spesso le regole ricavate sono valide solo all'interno di uno specifico dataset e non sono generalizzabili in più contesti.
- **Elevato costo computazionale:** nonostante siano in grado di gestire dataset di grandi dimensioni, gli algoritmi di mining delle regole associative, necessitano comunque di un numero crescente di capacità computazionale man mano che cresce anche il numero di transazioni,
- **Necessità di soglie minime:** i parametri fondamentali necessitano di soglie minime per poter essere utilizzati correttamente negli algoritmi, ciò verrà spiegato più dettagliatamente nei paragrafi successivi.

1.1.3.1 Algoritmo Apriori

L'algoritmo Apriori è un approccio iterativo in cui, a partire da insiemi frequenti di dimensione k uguale ad 1, si procede passo dopo passo alla ricerca di insiemi frequenti di dimensione $k+1$ utilizzando quelli di dimensione k trovati al passo precedente.

Le fondamenta di questo algoritmo risiedono nella *proprietà Apriori*[1], secondo la quale tutti i sottoinsiemi non vuoti di un insieme frequente devono essere a loro volta frequenti. Grazie ad essa, l'algoritmo stesso assume quindi che se un insieme dovesse risultare infrequente allora anche tutti i suoi superinsiemi lo saranno.

Prima di passare ad analizzare quali siano i passaggi che compongono la messa in pratica dell'algoritmo è necessario definire i valori di soglia minima dei due parametri fondamentali: supporto e confidenza. Mentre il supporto minimo andrà ad indicare il minimo numero di transazioni all'interno delle quali deve comparire un insieme per poter essere considerato frequente, la confidenza minima indica il minimo valore di "forza" che le regole devono avere per essere prese in considerazione (una regola con confidenza 0.65 indica che dato l'evento antecedente, esiste il 65% di probabilità che si verifichi l'evento conseguente).

Vediamo adesso quali sono i passaggi previsti da questo approccio:

1. Primo Passo: $k=1$.

- (a) Creazione di una tabella che associ ad ogni elemento del dataset il conteggio del proprio supporto.
- (b) Rimozione degli insiemi (di dimensione 1) ottenuti, aventi supporto inferiore al supporto minimo.

2. Secondo Passo: $k = k+1$.

La condizione da soddisfare per unire due itemset di dimensione k al fine di generarne uno di dimensione $k+1$, è che i due insiemi abbiano $k-2$ elementi in comune.

- (a) Si generano tutte le possibili combinazioni tra itemset di dimensione k che sono sopravvissuti alla selezione del passo precedente e si associa ad ognuno di essi il conteggio del proprio supporto ricavato contando il numero di transazioni in cui compare l'insieme. Questa fase è detta *fase di join*.
- (b) Rimozione degli insiemi ottenuti, aventi supporto inferiore al supporto minimo.

3. Passo N.

Si prosegue incrementando la dimensione k degli insiemi e ripetendo il passo 2 fino a che, al passo N , si hanno solo itemset con supporto inferiore al supporto minimo e quindi non frequenti. Ciò significa che gli insiemi frequenti di dimensione massima sono quelli trovati al passo $N-1$.

4. Generazione di regole associative forti.

Dato un insieme frequente di dimensione 3 si ricavano tutte le possibili regole associative ricombinando gli elementi dell'itemset e per ognuna di esse si calcola il valore di confidenza. A questo punto risulta essenziale la presenza di un valore di soglia minima in modo da poter discriminare le associazioni forti da quelle deboli da scartare. Alcuni esempi di regole (con confidenze casuali) derivabili dall'insieme $\{d1, d2, d3\}$ sono:

- $\{d1, d2\} \rightarrow \{d3\}$, confidenza = 50%
- $\{d1, d3\} \rightarrow \{d2\}$, confidenza = 50%
- $\{d1\} \rightarrow \{d2, d3\}$, confidenza = 30%
- $\{d2\} \rightarrow \{d1, d3\}$, confidenza = 25%

In questo caso, supponendo una soglia di confidenza minima del 50%, solo le regole 1 e 2 sarebbero da considerare forti.

Limitazioni: Questo algoritmo può risultare lento soprattutto nel caso in cui venga utilizzato per analizzare dataset di grandi dimensioni, a causa dell'elevato numero di insiemi che vengono creati (tutte le combinazioni di elementi) prima che avvenga la selezione in base al valore di supporto.

Capitolo 2

Interpretazione delle emozioni

Lo studio delle emozioni umane è da molto tempo uno dei campi di maggior interesse della psicologia. Molti psicologi si sono cimentati in questa ricerca imbattendosi però in non pochi ostacoli. Primo limite su tutti, l'eterogeneità degli esseri umani pone dei paletti non trascurabili alla realizzazione di una generica e assoluta interpretazione delle emozioni umane; si pensi ad esempio a come la differenza dei gusti personali faccia sì che persone diverse, provino la stessa emozione come reazione a stimoli diversi (o cambiando punto di vista: come uno stesso stimolo causi reazioni diverse a persone con gusti diversi). I gusti personali non sono l'unico aspetto dell'eterogeneità umana ad essere d'intralcio, ma anche: le esperienze vissute da una persona, le quali possono averla resa insensibile a certi input o estremamente sensibile ad altri, l'ambiente socio-culturale in cui essa si trova immersa o perfino l'età, che con i cambiamenti psicofisici che comporta con il proprio avanzare, può finire per invalidare eventuali regole emotive generiche e universali. Un secondo limite derivante in parte dal primo è quello della raccolta dati; risulta infatti estremamente difficile riuscire ad identificare con estrema precisione l'emozione suscitata in una persona se non attraverso l'utilizzo di sofisticati sensori in grado di captare ed analizzare le risposte fisiologiche del corpo umano ad un determinato stimolo, in quanto molto spesso persino colui che ha ricevuto direttamente l'input e ha vissuto in prima persona l'emozione non è in grado di assegnare valori specifici a ciò che ha provato.

2.1 James Russell e la Classificazione delle emozioni

Proprio perché sono molti i fattori di cui tenere conto quando si va a studiare lo spettro emotivo, esistono vari tipi di approccio per quanto riguarda la catalogazione delle emozioni, ognuno con i suoi vantaggi e svantaggi. Al giorno d'oggi non ne è stato ancora definito uno universalmente riconosciuto. Vediamo quali sono i principali:

- approccio dimensionale
- approccio categoriale: basato su associazioni tra segnali come espressioni facciali e determinate emozioni
- approccio neurologico: basato sull'attività cerebrale, come impulsi elettrici o identificazione delle aree attivate in risposta a diversi input

- approccio linguistico: basato su parole a cui sono state associate determinate emozioni
- approccio fisiologico: basato sulle risposte fisiologiche agli stimoli tramite sensori (si misurano parametri come: battito cardiaco, pressione sanguigna, conduttanza cutanea e altri ancora)
- approccio esperienziale: basato sul resoconto soggettivo delle emozioni

2.1.0.1 Approccio dimensionale

In questo tipo di approccio, come suggerisce il nome, l'obiettivo è mappare le emozioni su un piano mono-, bi- o tri-dimensionale, dove le dimensioni variano a seconda del modello scelto. Come si può dedurre facilmente anche solo procedendo per esclusione tra i vari approcci presentati nel paragrafo precedente, questo approccio risulta essere il più semplice da applicare oltre che il più versatile; si pensi ad esempio alla raccolta dati in forma remota ad esempio online. Entrambi i metodi di valutazione principali appartenenti a questo ramo della catalogazione delle emozioni, devono la propria esistenza allo psicologo statunitense James A. Russell, il quale ha contribuito alla realizzazione di un primo metodo chiamato *Modello di Russell-Mehrabian* o *PAD* e successivamente ha sviluppato un proprio metodo chiamato *Modello Circomplesso dell'Affetto*.

2.1.1 Modello PAD

Pubblicato tra il 1974 e il 1978, nato dalla collaborazione dei due psicologi James A. Russell e Anthony P. Mehrabian, il modello PAD vuole essere un'evoluzione del modello monodimensionale proposto qualche anno prima dal singolo Mehrabian. Questo modello predecessore, prevedeva la classificazione delle emozioni in base ad un unico parametro: *arousal* (inteso come "intensità"), rendendo poco distinguibili tra loro emozioni diverse a parità di potenza. Ecco quindi che nel modello successore vengono aggiunte altre due dimensioni: *pleasure* (dall'inglese "piacere", indicato come *valence*) e *dominance* (dall'inglese "dominanza") le quali indicano rispettivamente: la positività dell'emozione e la capacità di controllare la stessa. La tecnica di classificazione ottenuta è di tipo tridimensionale e permette di collocare le emozioni in un piano definito dai tre parametri i quali avevano un valore compreso tra 1 e 9.

2.1.2 Modello Circomplesso dell'Affetto

Pochi anni dopo la diffusione del modello PAD, nel 1980, James Russell pubblica un articolo[3] nel quale espone un nuovo modello rivoluzionario che ha lo scopo di migliorare il predecessore PAD: il Modello Circomplesso dell'Affetto.

L'idea principale di questa nuova proposta, è quella di passare dall'utilizzo di tre parametri (Pleasure, Arousal, Dominance) a utilizzarne otto (anche se i significati associati sono sette:

1. Pleasure (a cui ci riferiremo come Valence): indica se un'esperienza è percepita come positiva o negativa.

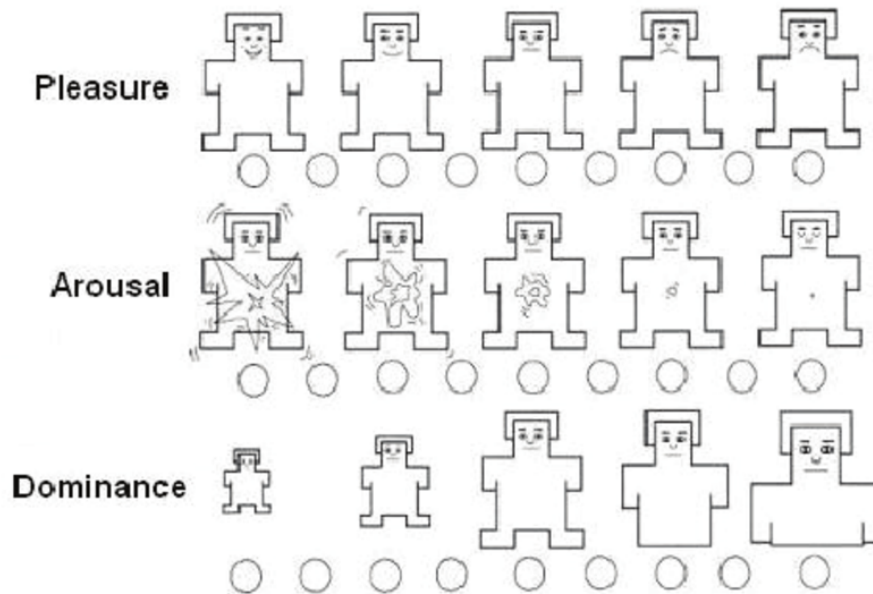


Figura 2.1: Scala di valutazione PAD: S.A.M.

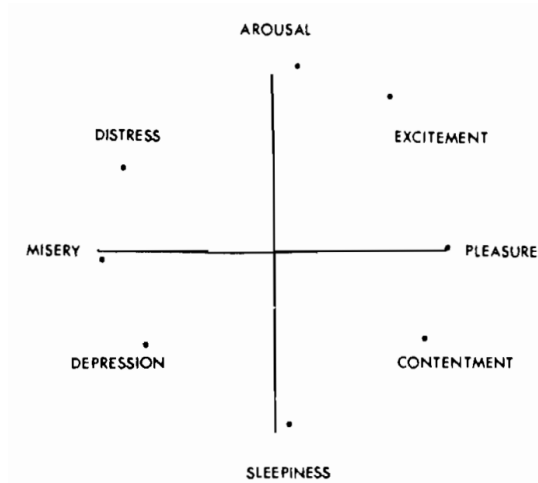
2. Arousal: grado di stimolazione provocato da un'esperienza.
3. Excitement: livello di entusiasmo percepito.
4. Contentment: grado di soddisfazione provocato da un'esperienza.
5. Sleepiness: livello di noia o sonnolenza.
6. Depression e Misery: grado di disperazione percepito.
7. Distress: livello di disagio o sofferenza emotiva sperimentato.

Come è facile immaginare, rappresentare un'emozione su otto dimensioni non è qualcosa di realizzabile, ecco quindi che l'intuizione di Russell entra in gioco: egli si rende conto del fatto che tutti e otto i parametri sono esprimibili in funzione di due valori fondamentali, Valence e Arousal, e che quindi il modello possa essere rappresentato su un piano bidimensionale. In particolare, le emozioni vengono disposte in modo circolare su questo piano dove, quelle aventi valori di valence ed arousal simili, saranno collocate in prossimità l'una dell'altra. Questa rappresentazione permette di dividere il piano in sezioni, ognuna delle quali è riconducibile ad un insieme di emozioni.

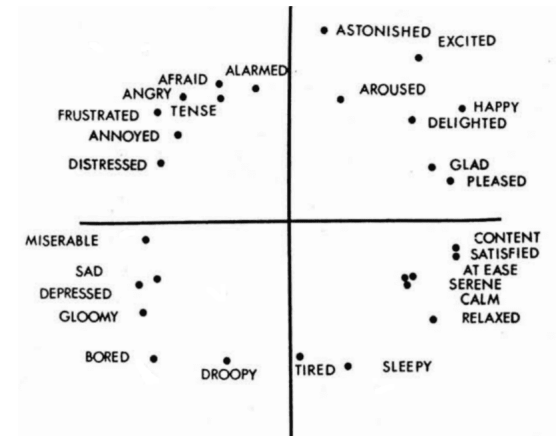
Ad oggi questo metodo risulta essere il più utilizzato tra quelli ad approccio dimensionale, grazie anche alla sua semplicità nella raccolta dati e nella loro successiva interpretazione.

2.1.2.1 Affective Slider

Poiché, come spiegato nel paragrafo precedente, secondo il modello circomplesso è possibile rappresentare tutte le emozioni basandosi solamente sui valori di valence e arousal, anche il meccanismo di raccolta dati varia rispetto a quello del modello PAD.



(a) Figura 2.2



(b) Figura 2.3

Figura 2.2: Collocazione degli otto parametri sul piano Valence(Pleasure)-Arousal

Figura 2.3: Collocazione delle emozioni sul piano Valence-Arousal

Mentre nell'approccio precedente si utilizzava il lo schema di valutazione denominato S.A.M. (Self-Assessment Manikin), con il nuovo approccio si preferisce l'utilizzo dell'Affective Slider, un test basato sull'assegnamento di un valore all'intensità e alla piacevolezza dell'emozione provata, spostando un cursore che scorre lungo un asse orizzontale. Viene così superato il limite imposto dal metodo di valutazione S.A.M., il quale richiedeva valori obbligatoriamente interi, permettendo adesso di mappare emozioni in un qualsiasi punto del piano.

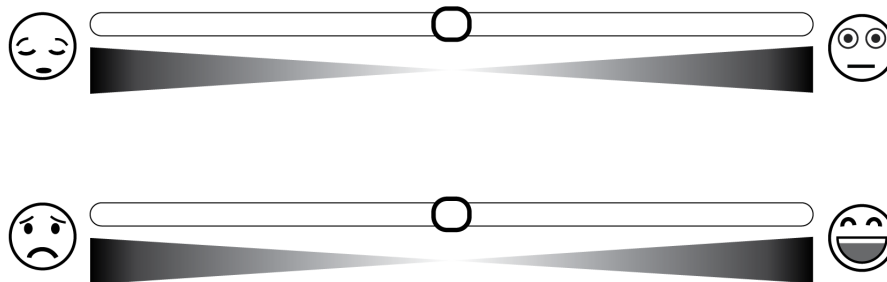


Figura 2.4: Affective Slider

il primo slider rappresenta valori di Arousal,
mentre il secondo quelli di Valence

2.2 Limitazioni modelli ad approccio dimensionale

Con i due modelli spiegati precedentemente, gli psicologi Russell e Mahrabian hanno cercato di superare uno dei tanti ostacoli che si oppongono sulla via della perfetta analisi emotiva, quello della classificazione delle emozioni. Nel perseguire questo obiettivo, è stata posta un'ipotesi fondamentale che è forse utopistica: i modelli funzionano correttamente se le persone assegnano valori corretti all'emozione percepita. Ecco che quindi si incontra subito un altro ostacolo, come assicurarsi che le persone esprimano correttamente ciò che provano e non assegnino valori casuali?

Capitolo 3

Affective Tagging e Data Collection

Nel capitolo precedente è stata descritta una limitazione importante dei modelli ad approccio dimensionale più comuni, per la quale sono state proposte negli anni un buon numero di possibili soluzioni. Durante la loro ricerca, il Dr. Pistolesi e l'Ing. Baldassini si sono imbattuti in questo problema e hanno deciso di evidenziare l'inattendibilità dei modelli di Russell e Mehrabian tramite la creazione di un nuovo metodo di valutazione.

3.1 Affective Tagging

Oggigiorno, con il frenetico utilizzo che la società fa degli smartphone e in particolare dei social media, le persone sono state costrette ad adattarsi e trovare un modo di esprimere e comunicare le proprie emozioni in maniera chiara, diretta e soprattutto da remoto: le emoji.

Da questa considerazione nasce l'idea di utilizzare le emoji come strumento di raccolta dati, data la familiarità che la maggior parte delle persone ha con esse, per poi confrontare i risultati ottenuti con quelli ricavati dai modelli dimensionali più diffusi.

3.1.1 Il test

Il test proposto dall'affective tagging prevede la visione di venti video, al termine di ognuno dei quali, viene richiesto all'utente di esprimere l'emozione percepita attraverso tre scale di valutazione:

1. Selezione delle emoji.
2. Affective Slider.
3. S.A.M.

I video proposti, estratti di film appartenenti ai generi più disparati (da commedia a horror, da drammatico a splatter), hanno una durata media di 2.5 minuti cadauno e mostrano in particolare, per ogni film, la scena che più rappresenta il genere cinematografico di appartenenza.

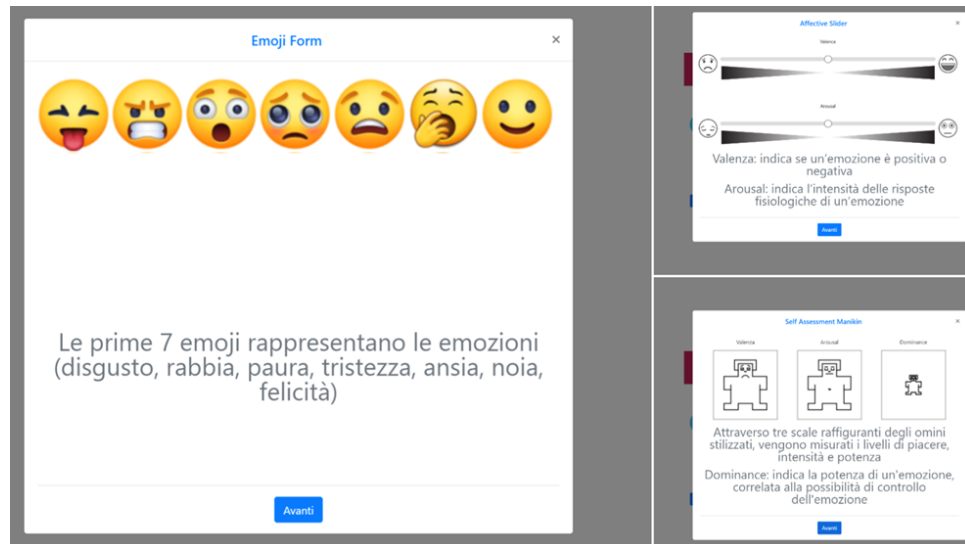


Figura 3.1: Tutorial mostrato ad inizio test

3.1.2 Il confronto

Ottenuti i valori nelle tre diverse scale di giudizio, le emoji selezionate, vengono collocate alle coordinate indicate dai valori di Valence e Arousal, ma come capire se i risultati siano coerenti con il piano di Russell?

A questo scopo, è stato creato un piano di Russell rivisitato, il quale assegna a determinate zone, le emoji più che le emozioni. Grazie a ciò è possibile valutare direttamente se la scelta effettuata dall'utente sia stata collocata nella corretta area.

3.2 Il mio contributo

Per quanto riguarda il mio coinvolgimento, l'obiettivo principale era quello di sottoporre al test quante più persone possibile, di età e sesso biologico più disparati, così da concentrare la mia attività di studio sulla ricerca di eventuali pattern.

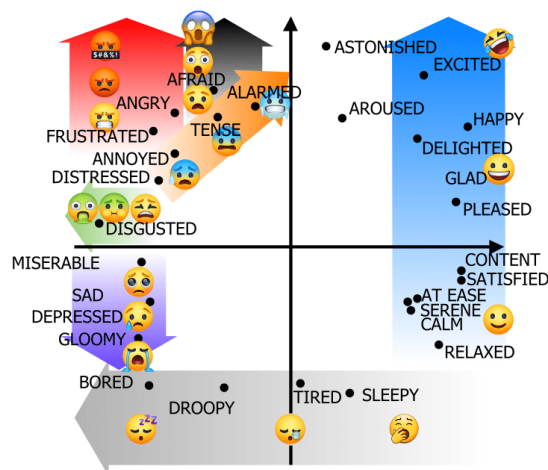


Figura 3.2: Piano di Russell rivisitato in ottica emoji

Capitolo 4

Analisi dei Dati e Conclusioni

Per l'analisi dei dati, ho scelto di utilizzare il linguaggio **python**. La mia scelta è stata dettata dal fatto che questo linguaggio e le sue numerose librerie si prestano in modo ideale alla gestione e analisi dei dati. Le librerie che ho utilizzato sono:

- **pandas**: fornisce la struttura tabellare **Dataframe**, la quale permette di lavorare con il dataset in analisi permettendo di manipolarne i dati. All'interno della libreria sono inoltre implementate funzioni che permettono di leggere dati da varie tipologie di file; un esempio è la funzione che ho utilizzato per leggere il dataset da un file CSV.
- **matplotlib**: è una delle librerie più popolari per la creazione di grafici. Nel mio caso è stata utilizzata per creare istogrammi rappresentanti la distribuzione delle persone sottoposte al test, nelle varie fasce d'età.
- **mlxtend**: questa libreria contiene l'implementazione di due funzioni essenziali al mio studio: *apriori*, la quale implementa la fase di ricerca di itemset frequenti, e *association_rules* che invece si occupa della ricerca di regole associative tra gli itemset frequenti.

4.1 Struttura dataset

Il dataset su cui ho lavorato è composto, al momento, da 3528 record dove ognuno di essi rappresenta l'insieme di valori selezionati, nei tre test, da un utente, dopo aver visionato un determinato video. Delle undici colonne di cui è composto (fig. 4.1), ho tenuto conto solo dei valori di: Age, Video, Emotion, Level, AS-Valence, AS-Arousal, SAM-Valence, SAM-Arousal. Per quanto riguarda i valori di AS-Valence e AS-Arousal, essi sono numeri decimali compresi nell'intervallo $[0.1, 0.9]$, mentre quelli di SAM-Valence e SAM-Arousal sono interi appartenenti all'intervallo $[1, 9]$. Il campo "Emotion" ha come valore una stringa appartenente all'insieme {anger, anxiety, boredom, disgust, fear, happiness, sadness} in base all'emoji scelta, mentre l'intensità è indicata dal campo "Level" avente valore nell'intervallo intero $[1, 3]$.

ID	Gender	Age	Video	Emotion	Level	AS-Valence	AS-Arousal	SAM-Valence	SAM-Arousal	SAM-Dominance
2	male	25	Crash-Contatto_fisico	anger	2	0.31	0.31	3	7	4
3	male	26	Crash-Contatto_fisico	anger	2	0.16	0.8	3	6	9
11	male	25	Gandhi	anger	2	0.26	0.26	3	7	7
12	female	27	Gandhi	boredom	1	0.5	0.5	5	5	5
29	male	25	Gli_Intoccabili-Stazione	anxiety	1	0.5	0.5	5	5	5
30	male	24	Gli_Intoccabili-Stazione	anxiety	3	0.33	0.9	4	9	2
46	male	27	Harry_ti_presento_Sally	happiness	3	1.25	1.22	9	1	9
47	female	21	Harry_ti_presento_Sally	happiness	3	0.5	0.5	9	5	5
18	female	26	IT_1-Trailer	anxiety	2	0.4	1.01	5	3	6
19	male	24	IT_1-Trailer	anxiety	3	1.15	0.57	5	7	8

Figura 4.1: Struttura dataset

4.2 Analisi Dataset

Il mio studio del dataset è iniziato con la ricerca di range, in cui suddividere le età dei partecipanti al test. Ho poi proseguito definendo dei valori corretti per Valence e Arousal sia nel test dell'Affective Slider sia nel test SAM, affinché l'emoji selezionata sia collocata nella corretta area del piano di Russell rivisitato. In una terza fase ho proceduto all'applicazione dell'algoritmo apriori e alla ricerca di regole associative, per poi analizzarle e trarre le conclusioni che saranno esposte poco più avanti.

4.2.1 Suddivisione utenti in fasce d'età

Le fasce di età che ho preso in considerazione come possibili candidati, sono tutti quelli appartenenti all'intervallo di interi $[5, 12]$ e, per ognuno di essi, ho ricavato un istogramma che rappresentasse graficamente la distribuzione di record e persone in ogni range. L'obiettivo era quello di trovare un candidato di dimensione tale da poter essere allo stesso tempo preciso (raggruppare in unica fascia persone di venti anni e di quaranta, potrebbe portare alla perdita di informazioni) ma anche abbastanza grande da includere un numero minimo di persone per fascia. Analizzando le varie possibilità, ho reputato che utilizzare intervalli temporali di 8 o 9 anni potesse essere il giusto compromesso tra precisione e distribuzione. Ho infine optato per raggruppamenti ad ottennio in quanto reputo che tra i due parametri, la precisione sia quello più importante e quindi a parità di condizioni il valore inferiore convenga.

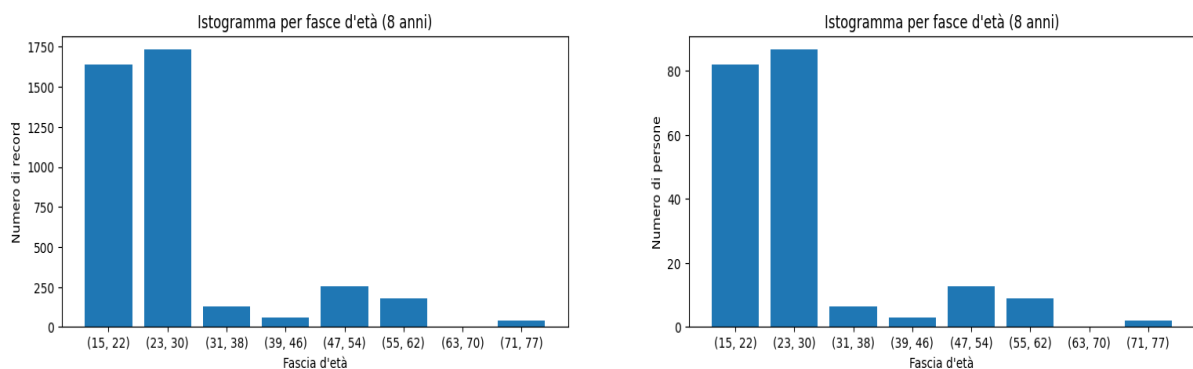


Figura 4.2: Istogramma intervalli da 8 anni

4.2.2 Decisione valori adeguati

La prima suddivisione che ho effettuato è stata quella dei video; ho infatti assegnato ogni filmato ad un'emozione così da poter discriminare se la scelta dell'emoji da parte di un utente sia corretta o meno. Tale associazione non deriva da un mio giudizio dei video ma da uno schema di "expected emotion" ricevuto dai miei relatori.

```

1 correct_emotion = {
2     #Happiness
3     'Laereo_piu_pazzo_del_mondo'      : {'Emotion': 'happiness'},
4     'Harry_ti_presento_Sally -Orgasmo' : {'Emotion': 'happiness'},
5     'Lassu_qualcuno_e_impazzito'      : {'Emotion': 'happiness'},
6
7     #Sadness
8     'Papa_ho_trovato_un_amico'        : {'Emotion': 'sadness'},
9     'Il_sapore_della_vittoria'        : {'Emotion': 'sadness'},
10    'The_Prestige -Morte_di_Julia'    : {'Emotion': 'sadness'},
11    'La_sottile_linea_rossa'          : {'Emotion': 'sadness'},
12
13    #Anger
14    'Gandhi'                          : {'Emotion': 'anger'},
15    'La_mia_guardia_del_corpo'         : {'Emotion': 'anger'},
16    'Crash-Contatto_fisico'           : {'Emotion': 'anger'},
17
18    #Fear
19    'Il_cigno_nero'                  : {'Emotion': 'fear'},
20    'Mulholland_Drive'               : {'Emotion': 'fear'},
21    'Silent_Hill'                    : {'Emotion': 'fear'},
22
23    #Anxiety
24    'IT_1-Trailer'                   : {'Emotion': 'anxiety'},
25    'IT_2-Trailer'                   : {'Emotion': 'anxiety'},
26    'Gli_Intoccabili -Stazione'      : {'Emotion': 'anxiety'},
27    'The_blair_witch_project'        : {'Emotion': 'anxiety'},
28
29    #Disgust
30    'Lesorcista'                     : {'Emotion': 'disgust'},
31    'The_Hannibal'                   : {'Emotion': 'disgust'},
32    'OldBoy'                         : {'Emotion': 'disgust'},
33 }
```

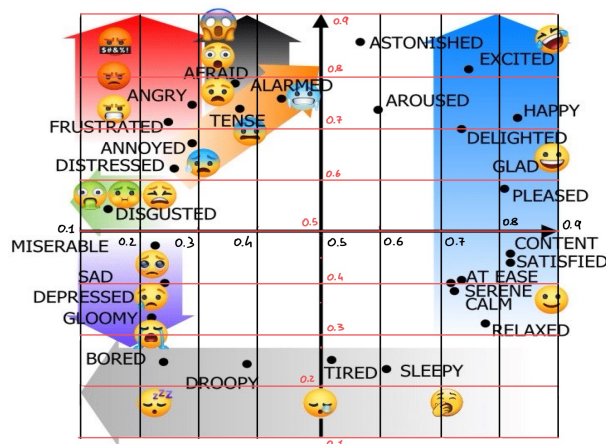


Figura 4.3: Suddivisione piano di Russell

Una volta associati video ed emoji, ho provveduto ad abbinare alle tuple (emoji, livello) i rispettivi intervalli di correttezza di Valence e Arousal nei due sistemi di valutazione considerati. Per determinare i vari range ho suddiviso il piano di Russell in otto intervalli (Fig 4.3) e considerato un leggero margine di errore per ogni emoji.

```

1 correct_intervals = {
2     #anger
3     ('anger', 1): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.6, 0.80],
4                   'SAM-Valence': [1, 3], 'SAM-Arousal': [6, 8]},
5     ('anger', 2): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.7, 0.85],
6                   'SAM-Valence': [1, 3], 'SAM-Arousal': [7, 8]},
7     ('anger', 3): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.8, 0.90],
8                   'SAM-Valence': [1, 3], 'SAM-Arousal': [8, 9]},
9
10    #anxiety
11    ('anxiety', 1): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.5, 0.7],
12                    'SAM-Valence': [1, 3], 'SAM-Arousal': [5, 7]},
13    ('anxiety', 2): {'AS-Valence': [0.2, 0.45], 'AS-Arousal': [0.6, 0.8],
14                    'SAM-Valence': [2, 4], 'SAM-Arousal': [6, 8]},
15    ('anxiety', 3): {'AS-Valence': [0.35, 0.6], 'AS-Arousal': [0.7, 0.9],
16                    'SAM-Valence': [3, 6], 'SAM-Arousal': [7, 9]},
17
18    #boredom
19    ('boredom', 1): {'AS-Valence': [0.6, 0.8], 'AS-Arousal': [0.1, 0.3],
20                    'SAM-Valence': [6, 8], 'SAM-Arousal': [1, 3]},
21    ('boredom', 2): {'AS-Valence': [0.3, 0.6], 'AS-Arousal': [0.1, 0.3],
22                    'SAM-Valence': [3, 6], 'SAM-Arousal': [1, 3]},
23    ('boredom', 3): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.1, 0.3],
24                    'SAM-Valence': [1, 3], 'SAM-Arousal': [1, 3]},
25
26    #disgust
27    ('disgust', 1): {'AS-Valence': [0.2, 0.4], 'AS-Arousal': [0.5, 0.6],
28                    'SAM-Valence': [2, 4], 'SAM-Arousal': [5, 6]},
29    ('disgust', 2): {'AS-Valence': [0.15, 0.3], 'AS-Arousal': [0.5, 0.6],
30                    'SAM-Valence': [2, 3], 'SAM-Arousal': [5, 6]},
31    ('disgust', 3): {'AS-Valence': [0.1, 0.2], 'AS-Arousal': [0.5, 0.6],
32                    'SAM-Valence': [1, 2], 'SAM-Arousal': [5, 6]},
33
34    #fear
35    ('fear', 1): {'AS-Valence': [0.3, 0.5], 'AS-Arousal': [0.60, 0.80],
36                 'SAM-Valence': [3, 5], 'SAM-Arousal': [6, 8]},
37    ('fear', 2): {'AS-Valence': [0.3, 0.5], 'AS-Arousal': [0.70, 0.85],
38                 'SAM-Valence': [3, 5], 'SAM-Arousal': [7, 8]},
39    ('fear', 3): {'AS-Valence': [0.3, 0.5], 'AS-Arousal': [0.80, 0.90],
40                 'SAM-Valence': [3, 5], 'SAM-Arousal': [8, 9]},
41
42    #happiness
43    ('happiness', 1): {'AS-Valence': [0.7, 0.9], 'AS-Arousal': [0.3, 0.5],
44                      'SAM-Valence': [7, 9], 'SAM-Arousal': [3, 5]},
45    ('happiness', 2): {'AS-Valence': [0.7, 0.9], 'AS-Arousal': [0.5, 0.7],
46                      'SAM-Valence': [7, 9], 'SAM-Arousal': [5, 7]},
47    ('happiness', 3): {'AS-Valence': [0.7, 0.9], 'AS-Arousal': [0.7, 0.9],
48                      'SAM-Valence': [7, 9], 'SAM-Arousal': [7, 9]},
49
50    #sadness
51    ('sadness', 1): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.4, 0.6],
52                    'SAM-Valence': [1, 3], 'SAM-Arousal': [4, 6]},
53    ('sadness', 2): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.3, 0.5],
54                    'SAM-Valence': [1, 3], 'SAM-Arousal': [3, 5]},
55    ('sadness', 3): {'AS-Valence': [0.1, 0.3], 'AS-Arousal': [0.1, 0.3],
56                    'SAM-Valence': [1, 3], 'SAM-Arousal': [1, 3]},
57 }

```


4.2.3 Applicazione algoritmo Apriori

Una volta organizzati i dati e definiti gli intervalli, ho elaborato un programma che, a seconda del metodo di self-assessment da prendere in esame, in ordine esegue le seguenti operazioni:

1. ricava una lista dei diversi titoli di video presenti nel dataset,
2. crea due colonne booleane basate sul metodo di valutazione (emoji, AS, SAM) passato nel campo 'method':
 - **correct_values:** contiene *True* solo in corrispondenza dei record che hanno associato i giusti valori ((Emoji, livello) o (AS-Valence, AS-Arousal) o (SAM-Valence, SAM-Arousal) a seconda del metodo scelto)
 - **incorrect_values:** come la precedente ma questa volta *True* è assegnato ai record dove i valori sono errati
3. itera sulla lista di video ricavata al punto 1 e per ognuno di essi:
 - (a) seleziona i soli record associati a quel video,
 - (b) crea una colonna contenente per ogni record l'intervallo di età di appartenenza dell'utente (**age_interval**)
 - (c) effettua codifica a caldo delle tre colonne interessate dall'algoritmo apriori (age_interval, correct_values, incorrect_values),
 - (d) applica algoritmo apriori con valore di supporto minimo dello 1% in modo da poter tenere conto anche delle fasce di età poco densamente popolate. Segue poi la ricerca di regole associative che restituisce solo regole con valore di lift ≥ 1.01 e confidenza minima del 98
 - (e) filtra le regole selezionando unicamente quelle che hanno come antecedente un range di età e le stampa.

```
1 import pandas as pd
2 from mlxtend.frequent_patterns import apriori
3 from mlxtend.frequent_patterns import association_rules
4 import utility
5
6 data = pd.read_csv('dataset.csv')
7
8 #salvo in una lista l'elenco distinct dei titoli dei video
9 videos_distinct = data['video'].unique().tolist()
10
11 #creo due colonne booleane che contengono esito comparazione tra valore
    assegnato e valore atteso
12 data = utility.apply_checks(data, method)
13
14 for video in videos_distinct:
15     video_records = data[data['video'] == video].copy()
16
17     #creo colonna contenente per ogni record il range d'età di appartenenza
18     video_records['age_interval'] = pd.cut(video_records['age'], bins=utility
        .bins_age, labels=utility.labels_age, right=False)
```

```

19
20     #codifica a caldo delle colonne interessate dall'algoritmo
21     data_encoded = pd.get_dummies(video_records[['age_interval', '
correct_values', 'incorrect_values']], prefix='', prefix_sep='')
22
23     #applico algoritmo Apriori
24     rules = utility.find_association_rules(data_encoded)
25
26     #filtro le regole che hanno come antecedente un range di eta'
27     rules_with_age_interval = rules[rules['antecedents'].astype(str).str.
contains(r'\d{2}-\d{2}')]
28
29     #stampo regole
30     utility.print_association_rules(video, rules_with_age_interval)

```

Le funzioni che ho utilizzato sono definite nel file utility.py che allego di seguito:

```

1  import pandas as pd
2  from mlxtend.frequent_patterns import apriori
3  from mlxtend.frequent_patterns import association_rules
4
5  #suddivisione eta' in intervalli di 8 anni
6  bins_age = list(range(15, 80, 8)) #creo intervalli di 8 anni da 15 a 79
7  labels_age = ['{}-{}'.format(i, i + 7) for i in range(15, 72, 8)] #creo
    etichette intervalli da 15 a 72 (devono essere una in meno rispetto ai
    bins)
8
9  #funzione che confronta valore assegnato nella colonna emotion con quello
    corretto
10 def em_check_correct_values(row, corr):
11     video = row['video']
12     emotion = row['Emotion']
13     right_em = correct_emotion[video]['Emotion']
14     if corr:
15         if emotion == right_em:
16             return True
17     else:
18         if emotion != right_em:
19             return True
20     return False
21
22 #funzione che confronta valori assegnato nelle colonne Valence e Arousal del
    test desiderato, con il range corretto di valori
23 def check_correct_values(row, column1, column2, corr):
24     video = row['video']
25     valence = row[column1]
26     arousal = row[column2]
27     emotion = row['Emotion']
28     level = row['Level']
29     target = (emotion, level)
30
31     valence_range = correct_intervals[target][column1]
32     arousal_range = correct_intervals[target][column2]
33
34     if corr:
35         if valence_range[0] <= valence <= valence_range[1] and arousal_range
[0] <= arousal <= arousal_range[1]:
36             return True
37     else:
38         if valence_range[0] > valence or valence > valence_range[1] or
arousal_range[0] > arousal or arousal > arousal_range[1]:
39             return True
40     return False
41
42 #funzione che crea colonne correct_values e incorrect_values

```

```

43 def apply_checks(data, check_func):
44     if check_func == 'EM':
45         data['correct_values'] = data.apply(em_check_correct_values, args=(
True,)), axis=1)
46         data['incorrect_values'] = data.apply(em_check_correct_values, args=(
False,)), axis=1)
47     elif check_func == 'AS':
48         data['correct_values'] = data.apply(check_correct_values, args=('AS-
Valence', 'AS-Arousal', True), axis=1)
49         data['incorrect_values'] = data.apply(check_correct_values, args=('AS
-Valence', 'AS-Arousal', False), axis=1)
50     elif check_func == 'SAM':
51         data['correct_values'] = data.apply(check_correct_values, args=('SAM-
Valence', 'SAM-Arousal', True), axis=1)
52         data['incorrect_values'] = data.apply(check_correct_values, args=('
SAM-Valence', 'SAM-Arousal', False), axis=1)
53     return data
54
55
56 #funzione che applica algoritmo apriori
57 def find_association_rules(data, min_support=0.01, min_confidence=0.98,
min_lift=1.01):
58     frequent_itemsets = apriori(data, min_support=min_support, use_colnames=
True)
59     rules = association_rules(frequent_itemsets, metric="lift", min_threshold
=min_lift)
60     return rules[rules['confidence'] >= min_confidence]
61
62 #funzione di stampa
63 def print_association_rules(video, rules):
64     print(video + ':')
65     print(rules)
66     print('\n')

```

4.2.4 Regole associative

In un primo momento ho applicato l'algoritmo ai tre metodi cercando unicamente relazioni che legassero unicamente range d'età e scelta di valori corretti o scorretti, ottenendo le regole associative presentate nelle successive tabelle.

Tabella 4.1: Regole associative con Affective Slider

Film	Antec.	Conseq.	Support	Confidence	Lift
Crash-Contatto_fisico	(39-46)	incorrect	0.013	1.000	1.328
Gandhi	(39-46)	incorrect	0.014	1.000	1.159
Gandhi	(55-62)	incorrect	0.043	1.000	1.159
Harry_ti_presento_Sally	(31-38)	incorrect	0.037	1.000	1.484
IT_1-Trailer	(39-46)	incorrect	0.018	1.000	1.207
IT_1-Trailer	(55-62)	incorrect	0.053	1.000	1.207
IT_2-Trailer	(31-38)	incorrect	0.029	1.000	1.245
IT_2-Trailer	(39-46)	incorrect	0.013	1.000	1.245
Il_cigno_nero	(31-38)	incorrect	0.037	1.000	1.402
Il_sapore_della_vittoria	(31-38)	incorrect	0.035	1.000	1.117
Il_sapore_della_vittoria	(39-46)	incorrect	0.017	1.000	1.117

Continua nella pagina successiva

Tabella 4.1 – continuazione dalla pagina precedente

Film	Antec.	Conseq.	Support	Confidence	Lift
Il_sapore_della_vittoria	(55-62)	incorrect	0.052	1.000	1.117
La_mia_guardia_del_corpo	(31-38)	incorrect	0.033	1.000	1.200
La_mia_guardia_del_corpo	(39-46)	incorrect	0.017	1.000	1.200
La_sottile_linea_rossa	(31-38)	incorrect	0.036	1.000	1.072
La_sottile_linea_rossa	(39-46)	incorrect	0.016	1.000	1.072
La_sottile_linea_rossa	(55-62)	incorrect	0.047	1.000	1.072
Lassu_qualcuno_e_impazzito	(31-38)	incorrect	0.036	1.000	1.129
Lassu_qualcuno_e_impazzito	(39-46)	incorrect	0.018	1.000	1.129
OldBoy	(15-22)	incorrect	0.346	0.982	1.027
OldBoy	(31-38)	incorrect	0.038	1.000	1.046
OldBoy	(39-46)	incorrect	0.019	1.000	1.046
Papa_ho_trovato_un_amico	(39-46)	incorrect	0.015	1.000	1.170
Silent_Hill	(39-46)	incorrect	0.019	1.000	1.285
Silent_Hill	(55-62)	incorrect	0.057	1.000	1.285
The_Hannibal	(31-38)	incorrect	0.037	1.000	1.046
The_Hannibal	(39-46)	incorrect	0.019	1.000	1.046
The_Hannibal	(55-62)	incorrect	0.056	1.000	1.046
Fine della tabella					

Dall'analisi di questa prima tabella il primo dato che cattura l'attenzione è rappresentato proprio dalla colonna *Consequents*. Infatti, si può notare come tutte le regole indichino che se l'utente appartiene ad un determinato range di età, allora assegnerà alle proprie emozioni dei valori errati sull'Affective Slider. Questo fenomeno coinvolge praticamente solo gli utenti che hanno superato i 30 anni di età, infatti, sebbene si possa scovare una regola che lega ragazzi della fascia (15-22) all'errore nel caso di una scena triste, i valori di confidence e lift (i più bassi di tutta la tabella) ci indicano che questa regola è quella di minor forza tra quelle presentate. In particolare si può evincere che il maggior numero di errori viene commesso proprio dalle persone di età appartenente all'intervallo [31, 46].

Tabella 4.2: Regole associative con SAM

Film	Antec.	Conseq.	Support	Confidence	Lift
Crash-Contatto_fisico	(39-46)	incorrect	0.013	1.000	1.730
Harry_ti_presento_Sally	(39-46)	incorrect	0.016	1.000	1.272
IT_2-Trailer	(39-46)	incorrect	0.013	1.000	1.440
La_sottile_linea_rossa	(31-38)	incorrect	0.036	1.000	1.350
La_sottile_linea_rossa	(39-46)	incorrect	0.016	1.000	1.350
Laereo_piu_pazzo_del_mondo	(39-46)	incorrect	0.016	1.000	1.217
Laereo_piu_pazzo_del_mondo	(47-54)	incorrect	0.070	1.000	1.217
Lassu_qualcuno_e_impazzito	(55-62)	incorrect	0.054	1.000	1.161
Mulholland_Drive	(39-46)	incorrect	0.018	1.000	1.443
Continua nella pagina successiva					

Tabella 4.2 – continuazione dalla pagina precedente

Film	Antec.	Conseq.	Support	Confidence	Lift
Mulholland_Drive	(47-54)	incorrect	0.078	1.000	1.443
Mulholland_Drive	(55-62)	incorrect	0.054	1.000	1.443
OldBoy	(31-38)	incorrect	0.038	1.000	1.136
OldBoy	(39-46)	incorrect	0.019	1.000	1.136
Silent_Hill	(39-46)	incorrect	0.019	1.000	1.463
Silent_Hill	(55-62)	incorrect	0.057	1.000	1.463
The_Hannibal	(31-38)	incorrect	0.038	1.000	1.111
The_Hannibal	(55-62)	incorrect	0.056	1.000	1.111
The_Prestige-Morte_di_Julia	(39-46)	incorrect	0.017	1.000	1.496
Fine della tabella					

Studiando la tabella risultante dall'analisi dei risultati SAM, si giunge rapidamente ad una conclusione molto simile alla precedente. Infatti anche in questo caso, i range di età coinvolti negli errori sono appartenenti alla categoria over 30, con la fascia (39-46). Da notare è, che seppure di qualche punto decimale, in questa seconda analisi si riscontrano valori di lift maggiori; ciò indica una minor casualità della correlazione.

Tabella 4.3: Regole associative con Emoji

Film	Antec.	Conseq.	Support	Confidence	Lift
Crash-Contatto_fisico	(31-38)	correct	0.039	1.000	1.222
Crash-Contatto_fisico	(39-46)	correct	0.013	1.000	1.222
Harry_ti_presento_Sally	(39-46)	correct	0.016	1.000	1.273
Harry_ti_presento_Sally	(55-62)	correct	0.048	1.000	1.119
IT_2-Trailer	(39-46)	correct	0.013	1.000	1.441
Il_sapore_della_vittoria	(39-46)	correct	0.017	1.000	1.344
La_sottile_linea_rossa	(39-46)	correct	0.016	1.000	1.485
Lassu_qualcuno_e_impazzito	(39-46)	correct	0.018	1.000	1.627
Mulholland_Drive	(39-46)	incorrect	0.018	1.000	1.694
OldBoy	(39-46)	correct	0.019	1.000	1.420
OldBoy	(55-62)	correct	0.057	1.000	1.420
Papa_ho_trovato_un_amico	(31-38)	correct	0.035	1.000	1.075
Papa_ho_trovato_un_amico	(39-46)	correct	0.015	1.000	1.075
Papa_ho_trovato_un_amico	(47-54)	correct	0.065	1.000	1.075
Silent_Hill	(31-38)	incorrect	0.038	1.000	1.505
Silent_Hill	(39-46)	incorrect	0.019	1.000	1.505
The_Hannibal	(31-38)	correct	0.038	1.000	1.270
The_Prestige-Morte_di_Julia	(39-46)	incorrect	0.017	1.000	1.549
The_blair-witch_project	(31-38)	correct	0.039	1.000	2.153
Fine della tabella					

La tabella che riguarda il test tramite emoji è quella che contiene i dati più positivi tra le tre. Infatti, si nota una quasi totalità di regole che hanno come conseguenza la

corretta identificazione dell'emozione. Anche in questo caso i range di età appartengono tutti alla categoria over 30; regole che coinvolgano anche range minori iniziano a palesarsi al calare del valore di confidence minima che però deve rimanere elevato dato che è già stata applicata una forte riduzione di supporto minimo in modo da non trascurare le minoranze.

4.2.5 Risultati

Per la valutazione dei risultati finali verrà posta come ipotesi fondamentale quella per cui, l'emozione suscitata negli utenti sia quella prevista dal test e che quindi in caso di incongruenze sia il metodo valutativo a non essere stato in grado di esprimerla correttamente. Le regole che ho ottenuto sono definite **regole di nicchia**, si tratta cioè di relazioni nascoste all'interno del dataset a causa della loro bassa frequenza di manifestazione oppure a causa delle circostanze particolari che richiedono. Essendo regole di questo tipo, sono considerabili regole forti nonostante il valore di lift non superi mai di troppo il valore minimo affinché la regola sia definibile non casuale (1).

Cosa è possibile estrapolare da esse? Vediamo nel dettaglio:

- per quanto riguarda gli utenti appartenenti al range di età (31, 38), l'utilizzo di Affective Slider si dimostra la scelta peggiore. Infatti è vero che, nella tabella corrispondente, esiste una regola per almeno un video a categoria tale che questa fascia di età sia associata ad una valutazione errata. Al contrario, Il metodo SAM si dimostra migliore rispetto ad A.S. per le persone di questo gruppo; esistono infatti solo due regole che legano questa fascia ad una valutazione errata (in particolare si registra errore in corrispondenza di emozioni quali disgusto e tristezza). Tuttavia l'utilizzo di Emoji risulta essere il metodo migliore per questi utenti: non solo il numero di regole che li associa ad una scorretta valutazione è inferiore a quello di tutti gli altri metodi (soltanto una) ma sono anche presenti quattro relazioni con una corretta valutazione in corrispondenza di video della categoria disgusto, rabbia, ansia e tristezza.
- per gli utenti aventi età appartenente al range (39, 46), l'utilizzo di Affective Slider risulta essere estremamente inefficiente. Esiste infatti una regola che associa questa fascia all'errore per almeno un film a categoria di emozione mentre non esistono regole che la associno a valutazioni corrette. La stessa considerazione vale anche per il metodo SAM. Andando a considerare la valutazione tramite emoji, si può notare un importante miglioramento: la tabella relativa a questo metodo mostra la presenza di cinque regole che legano questa nicchia di utenti ad una corretta interpretazione e soltanto due regole con interpretazione errata.
- riguardo la fascia (47-54) si nota che la tabella di Affective Slider non contiene regole a riguardo, ciò può significare che non ci sono particolari pattern che li riguardano. Il metodo SAM si dimostra inefficiente nell'identificazione di paura e felicità, mentre tramite Emoji viene valutata in modo corretto la tristezza.
- per gli utenti della fascia (55-62) si osserva che, Affective Slider si afferma nuovamente peggior metodo data la presenza di una regola legata all'incorrettezza in

corrispondenza di almeno un video appartenente a cinque categorie diverse. SAM conferma di essere l'intermedio tra AS ed Emoji con tre regole legate ad una valutazione negativa e infine il metodo Emoji sembra essere valevole in particolare per identificare le emozioni agli estremi dell'asse orizzontale sul piano di Russell; si trovano regole che legano questa fascia ad una corretta valutazione nel caso di disgusto e di felicità.

Ho inoltre cercato altri possibili dettagli nascosti nel dataset; in primis ho cercato di identificare se esistessero, essendo queste regole di nicchia, delle soglie di supporto minimo oltre le quali l'algoritmo apriori non fosse più stato in grado di restituire regole associative.

- per la valutazione ad Emoji il valore massimo di supporto minimo per cui viene generata almeno una regola associativa è 0.06, corrispondente quindi al 6%.
- per il metodo SAM, invece, il massimo valore di supporto minimo tollerato è 0.07, equivalente al 7%.
- infine, per l'Affective Slider, il massimo valore è 0.3, cioè il 30%. Da ciò si deduce che tra tutte le regole generate quelle meno "di nicchia" appartengono a questo insieme.

Come ultimo tentativo nella ricerca di pattern nascosti ho deciso di aggiungere il genere degli utenti, tra i campi da far valutare all'algoritmo apriori. Mi sono concentrato in particolare sul metodo ad emoji che ha dimostrato migliori performance in precedenza e ciò che ho ottenuto è il contenuto della tabella sottostante dove, le righe evidenziate indicano regole nuove ottenute totalmente grazie all'aggiunta del parametro 'gender'.

Tabella 4.4: Regole associative con genere nel metodo Emoji

Film	Antec.	Conseq.	Support	Confidence	Lift
Crash-Contatto_fisico	(female, 39-46)	correct	0.013	1.000	1.222
Crash-Contatto_fisico	(male, 47-54)	correct	0.019	1.000	1.222
Crash-Contatto_fisico	(female, 55-62)	correct	0.026	1.000	1.222
Gandhi	(male, 47-54)	correct	0.014	1.000	1.835
Gandhi	(female, 31-38)	incorrect	0.009	1.000	2.198
Gli_Intoccabili-Stazione	(female, 31-38)	incorrect	0.011	1.000	1.945
Harry_ti_presento_Sally	(female, 31-38)	correct	0.011	1.000	1.120
Harry_ti_presento_Sally	(female, 39-46)	correct	0.016	1.000	1.120
IT_1-Trailer	(female, 31-38)	correct	0.012	1.000	2.485
IT_1-Trailer	(male, 47-54)	incorrect	0.018	1.000	1.673
IT_1-Trailer	(female, 55-62)	incorrect	0.024	1.000	1.673
IT_2-Trailer	(female, 39-46)	correct	0.013	1.000	2.213
Il_cigno_nero	(female, 31-38)	correct	0.012	1.000	2.412
Il_sapore_della_vittoria	(female, 39-46)	correct	0.017	1.000	1.344
Il_sapore_della_vittoria	(male, 47-54)	correct	0.017	1.000	1.344
Il_sapore_della_vittoria	(female, 55-62)	correct	0.023	1.000	1.344
La_mia_guardia_del_corpo	(male, 31-38)	correct	0.022	1.000	1.525

Continua nella pagina successiva

Tabella 4.4 – continuazione dalla pagina precedente

Film	Antec.	Conseq.	Support	Confidence	Lift
La_sottile_linea_rossa	(male, 31-38)	correct	0.026	1.000	1.485
La_sottile_linea_rossa	(female, 39-46)	correct	0.016	1.000	1.485
La_sottile_linea_rossa	(male, 47-54)	correct	0.016	1.000	1.485
La_sottile_linea_rossa	(female, 55-62)	correct	0.021	1.000	1.485
Laereo_piu_pazzo_del_mondo	(male, 31-38)	correct	0.022	1.000	1.259
Laereo_piu_pazzo_del_mondo	(male, 55-62)	correct	0.027	1.000	1.259
Laereo_piu_pazzo_del_mondo	(female, 31-38)	incorrect	0.011	1.000	4.868
Lassu_qualcuno_e_impazzito	(female, 39-46)	correct	0.018	1.000	1.627
Lassu_qualcuno_e_impazzito	(male, 47-54)	correct	0.018	1.000	1.627
Lesorcista	(female, 31-38)	correct	0.013	1.000	4.618
Lesorcista	(female, 39-46)	incorrect	0.019	1.000	1.276
Lesorcista	(male, 47-54)	incorrect	0.019	1.000	1.276
Lesorcista	(female, 55-62)	incorrect	0.025	1.000	1.276
Mulholland_Drive	(female, 31-38)	incorrect	0.012	1.000	1.694
Mulholland_Drive	(female, 39-46)	incorrect	0.018	1.000	1.694
OldBoy	(female, 31-38)	correct	0.013	1.000	1.420
OldBoy	(female, 39-46)	correct	0.019	1.000	1.420
Papa_ho_trovato_un_amico	(female, 23-30)	correct	0.085	1.000	1.075
Papa_ho_trovato_un_amico	(female, 39-46)	correct	0.015	1.000	1.075
Silent_Hill	(female, 39-46)	incorrect	0.019	1.000	1.505
The_Hannibal	(female, 55-62)	correct	0.025	1.000	1.270
The_Prestige-Morte_di_Julia	(female, 31-38)	incorrect	0.011	1.000	1.549
The_Prestige-Morte_di_Julia	(female, 39-46)	incorrect	0.017	1.000	1.549
The_Prestige-Morte_di_Julia	(male, 47-54)	incorrect	0.017	1.000	1.549
Fine della tabella					

4.3 Conclusioni

Da questo studio si possono evincere informazioni interessanti riguardo sia i metodi di self-assessment, sia i modi in cui le persone quantificano le proprie emozioni.

Basandosi su quanto dimostrato nell'analisi dei risultati al paragrafo precedente, si evincono due informazioni degne di nota: la prima è data dal fatto che gli utenti preferiscono usare, o per lo meno utilizzano più efficacemente, metodi di valutazione dimensionali che abbiano valori fissi tra cui scegliere. Ciò è evidenziato dal fatto che i due metodi con questo tipo di approccio (SAM ed Emoji) hanno sempre dimostrato esiti migliori rispetto ad Affective Slider. La seconda invece è data dalle prestazioni di gran lunga migliori fornite dall'innovativo metodo Emoji; esso sembra essere più intuitivo, non solo per i giovani che ormai sono abituati ad un uso quotidiano di questo modo di esprimere le emozioni, ma anche dai più grandi. Infine è importante tenere conto del fatto che l'età non è l'unico fattore da considerare quando si cerca di prevedere l'emozione che proverà una certa persona al vivere una determinata esperienza. Già solo considerando un parametro come il sesso biologico di nascita può portare a scoprire ulteriori relazioni utili ma nascoste.

4.3.1 Use Cases

Le informazioni ricavate da questo studio, se maggiormente sviluppate ad esempio facendo sostenere il test dell'Affective Tagging ad un numero sempre maggiore di utenti di età e genere più disparati, potrebbero portare ad una comprensione dello spettro emotivo tale da apportare migliorie a qualsiasi settore che abbia come destinatario un essere umano e quindi possa trarre beneficio dalla possibilità di prevedere l'impatto di un proprio prodotto sui clienti. In particolare reputo che le applicazioni più proficue sarebbero nell'ambito del marketing pubblicitario, dell'intrattenimento e nell'ambito informatico; proprio in quest'ultimo, riuscendo a prevedere le emozioni degli utenti, si potrebbe creare una user experience personalizzata di livello molto superiore rispetto a quelle moderne.

Ringraziamenti

Ringrazio i miei relatori Dr.Francesco Pistolesi e Ing.Michele Baldassini per avermi introdotto al Data Mining e alla loro ricerca sull'Affective Tagging e soprattutto per avermi assistito durante la conclusione del mio percorso triennale dimostrando molta pazienza.

Bibliografia

- [1] Mohityadav. Apriori algorithm. <https://www.geeksforgeeks.org/apriori-algorithm/>, 2022.
- [2] Mohityadav. Frequent item set in data set (association rule mining). <https://www.geeksforgeeks.org/frequent-item-set-in-data-set-association-rule-mining/>, 2023.
- [3] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980. doi:10.1037/h0077714.
- [4] Amazon Web Services. Che cos'è il data mining? <https://aws.amazon.com/it/what-is/data-mining/>.