UNIVERSITÀ DEGLI STUDI DI SALERNO

Progetto di Intelligenza Artificiale

Studente	**Matricola**
Scaparra Daniele Pio	0512116260
Fasolino Pietro	0512116473
Vitulano Antonio	0512116776

Link alla repository GitHub:
https://github.com/dscap02/EmotionsReleave

Documento di Analisi Dati

Indice

1 Introduzione

Questo documento descrive l'analisi del dataset GoEmotion. Verranno trattati i passi di comprensione dei dati, analisi e qualità, oltre all'individuazione delle variabili chiave per la costruzione del modello.

2 Data Understanding

2.1 Descrizione del Dataset

Sulla base degli obiettivi prefissati, ci concentriamo sull'analisi del dataset GoEmotion, una collezione di dati composta da 58.000 commenti prelevati da Reddit. Ogni commento è etichettato con una o più emozioni, per un totale di 28 emozioni (27 emozioni specifiche + una neutrale).

2.2 Struttura del Dataset

Il dataset contiene le seguenti colonne:

- **text:** Il testo del commento (con token mascherati come [NAME] o [RELIGION]).
- **id:** L'ID univoco del commento.
- **author:** L'utente Reddit autore del commento.
- **subreddit:** Il subreddit di appartenenza.
- **link_id:** L'ID del link del commento.
- **parent_id:** L'ID del commento padre.
- **created_utc:** Il timestamp del commento.
- **rater_id:** L'ID univoco dell'annotatore.
- **example_very_unclear:** Indica se l'esempio è stato marcato come poco chiaro.

3 Analisi dei Dati

3.1 Classificazione delle Emozioni

Il dataset comprende 28 emozioni:

- **12 emozioni positive:** Ammirazione, Divertimento, Entusiasmo, Gratitudine, Gioia, Amore, ecc.
- **11 emozioni negative:** Rabbia, Disgusto, Paura, Dolore, Tristezza, ecc.

- **4 emozioni ambigue:** Confusione, Curiosità, ecc.
- **1 emozione neutrale.**

Ogni emozione è rappresentata da un indice numerico. Ad esempio:

- **Ammirazione (0):** Provare qualcosa di impressionante o degno di rispetto.
- **Divertimento (1):** Trovare qualcosa di divertente o intrattenersi.
- **Rabbia (2):** Un forte sentimento di dispiacere o antagonismo.
- **Fastidio (3)**: Lieve rabbia o irritazione.
- **Approvazione (4)**: Avere o esprimere un'opinione favorevole.
- **Premura (5)**: Mostrare gentilezza e interesse per gli altri.
- **Confusione (6)**: Mancanza di comprensione, incertezza.
- **Curiosità (7)**: Un forte desiderio di conoscere o apprendere qualcosa.
- **Desiderio (8)**: Un forte sentimento di voler qualcosa o di sperare che accada.
- **Delusione (9)**: Tristezza o dispiacere causati dal non raggiungimento delle aspettative.
- **Disapprovazione (10)**: Avere o esprimere un'opinione sfavorevole.
- **Disgusto (11)**: Repulsione o forte disapprovazione verso qualcosa di sgradevole o offensivo.
- **Imbarazzo (12)**: Imbarazzo, vergogna o senso di disagio.
- **Entusiasmo (13)**: Sentimento di grande eccitazione e impazienza.
- **Paura (14)**: Essere spaventati o preoccupati.
- **Gratitudine (15)**: Un sentimento di riconoscenza e apprezzamento.
- **Dolore (16)**: Tristezza intensa, specialmente causata dalla perdita di qualcuno.
- **Gioia (17)**: Sentirsi felici e contenti.
- **Amore (18)**: Una forte emozione positiva di affetto e amore.
- **Nervosismo (19)**: Apprensione, preoccupazione o ansia.
- **Ottimismo (20)**: Speranza e fiducia riguardo al futuro.
- **Orgoglio (21)**: Piacere o soddisfazione per i propri successi o quelli di persone vicine.

- **Consapevolezza (22)**: Diventare consapevoli di qualcosa.
- **Sollievo (23)**: Rassicurazione e rilassamento dopo uno stato di ansia o stress.
- **Rimorso (24)**: Rimpianto o senso di colpa.
- **Tristezza (25)**: Dolore emotivo o sofferenza.
- **Sorpresa (26)**: Sentirsi stupiti o sorpresi da qualcosa di inaspettato.
- **Neutra (27)** Quando non provi nulla nei confronti dell'altra persona sia emozioni positive che negative.

3.2 Peculiarità del Dataset

Alcuni messaggi possono avere più emozioni associate, ad esempio:

- mess1 con emozione 0
- mess2 con emozioni 0,1
- mess3 con emozioni 0,1,4

Ulteriori dettagli statistici sono disponibili nella cartella **documentation** e nei file correlati.

4 Individuazione Variabili Chiave e Correlazioni

Le variabili chiave individuate sono:

- **Messaggio:** Il testo del commento.
- **Emozioni:** Le etichette emozionali associate al messaggio.

Altre variabili come **author** e **subreddit** hanno importanza secondaria per il nostro modello.

Sono stati rilevati problemi di correlazione tra emozioni. Alcune emozioni appaiono raramente e sono associate a combinazioni poco frequenti, aumentando il rischio di rumore.

5 Qualità dei Dati

In questa fase analizziamo eventuali problemi di qualità dei dati, con particolare attenzione a problematiche come la mancanza di dati,la ridondanza,il rumore e anche sulle correlazioni. Nel nostro caso, l'analisi del dataset di base ha mostrato che sono già state applicate operazioni di pulizia per risolvere i problemi di mancanza o ridondanza. Di conseguenza, il dataset risulta completo sotto i primi due aspetti, e non è stato necessario alcun ulteriore intervento.

Durante l'analisi dei dati,documentata apparte(emozioni.pdf), è emerso però un problema di rumore dei dati,cioè che alcune combinazioni di emozioni erano presenti in un numero estremamente ridotto di istanze. Queste combinazioni, costituite da emozioni principali associate a emozioni molto rare, possono essere considerate rumore nel dataset, poiché la loro scarsità non fornisce informazioni sufficienti per un apprendimento significativo da parte del modello.

Tuttavia, un altro problema rilevato durante l'analisi delle singole emozioni (riportata in un documento separato, emozioni.pdf) riguarda la presenza di istanze con target che includono più emozioni associate allo stesso messaggio. Questa condizione potrebbe rappresentare uno svantaggio nella fase di addestramento del modello, in quanto il modello potrebbe imparare da dati che combinano emozioni diverse. Ciò rischia di introdurre confusione durante la predizione finale, portando potenzialmente a risultati meno accurati.

6 Conclusione

Il dataset GoEmotion offre un'ampia varietà di emozioni, ma presenta alcune sfide, come il rumore nei dati e la complessità delle correlazioni emozionali. Tali questioni verranno affrontate nella fase successiva di preparazione dei dati.

A Appendice

Ulteriori analisi, grafici e statistiche si trovano nella cartella **Results**.