

[pair-code.github.io/datacardsplaybook](http://pair-code.github.io/datacardsplaybook)  
Data Cards Template (Beta) | Updated November 2022

This document is designed for use in Google Docs, Microsoft Office, Libre Office, or other similar text editing software. Find PDF and Markdown versions of this document in our Github repository or on our website.

| Dataset Name (Acronym) | Il dataset è composto da 5 colonne: "User" per i nomi utente, "Date" per le date, "Stars" per il numero di stelle assegnate dagli utenti, "Text" per le recensioni e "Title" per i titoli delle recensioni. Per la sentiment analysis, mi serviranno solo le colonne "Text" e "Stars". Ho 315 osservazioni nel dataset. |
| --- | --- |
| DATASET LINK | DATA CARD AUTHOR(S) |
| Provide a link to the dataset: | Select **one role per** Data Card Author:  (Usage Note: Select the most appropriate choice to describe the author's role in creating the Data Card.) |
| https://www.imdb.com/title/tt13210838/reviews/ | **Valentina Baldi :** (Owner ) |

| Authorship | | |
| --- | --- | --- |
| Publishers | | |
| PUBLISHING ORGANIZATION(S) | INDUSTRY TYPE(S) | CONTACT DETAIL(S) |
| Provide the names of the institution or organization responsible for publishing the dataset: | Select **all applicable** industry types to which the publishing organizations belong: | Provide publisher contact details: |
| Università LUMSA-Libera Università Maria Ss. Assunta | Academic - Tech | **Publishing POC:** <Valentina Baldi .>  **Affiliation:** < Università LUMSA.>  **Contact:** < info@**lumsa**.it>  **Website:** < https://www.imdb.com/.> |
| Dataset Owners | | |
| TEAM(S) | CONTACT DETAIL(S) | AUTHOR(S) |
| Provide the names of the groups or team(s) that own the dataset: | Provide pathways to contact dataset owners: | Provide the details of all authors associated with the dataset:  (Usage Note: Provide the affiliation and year if different from publishing institutions or multiple affiliations.) |
| Valentina Baldi | **Dataset Owner(s):** <Valentina Baldi>  **Affiliation:** <Università LUMSA>  **Contact:** <v.baldi@lumsastud.com> | Valentina Baldi, Owner |

###### 

###### 

| Dataset Overview | | |
| --- | --- | --- |
| DATA SUBJECT(S) | DATASET SNAPSHOT | CONTENT DESCRIPTION |
| Select **all applicable** subjects contained the dataset: | Provide a snapshot of the dataset:  (Use the additional notes to include relevant information, considerations, and links to table(s) with more detailed breakdowns.) | Provide a short description of the content in a datapoint. |
| Non-Sensitive Data about peopleData riguardo cosa piace o meno alle persone | | Size of Dataset | 307 KB | | --- | --- | | Number of Instances | 315 | | Number of Fields | 5 | | Labeled Classes | 123456 | | Number of Labels | 10 | | Average Labels Per Instance | 8 | | Il dataset è composto da 5 colonne: "User" per i nomi utente, "Date" per le date, "Stars" per il numero di stelle assegnate dagli utenti, "Text" per le recensioni e "Title" per i titoli delle recensioni. Per la sentiment analysis, mi serviranno solo le colonne "Text" e "Stars". Ho 315 osservazioni nel dataset.  **Additional Notes:** link dove ho preso I dati: https://www.imdb.com/title/tt13210838/reviews/  ?ref\_=tt\_ov\_rt |
| DESCRIPTIVE STATISTICS | | |
| Provide basic descriptive statistics for each field.Use additional notes to capture any other relevant information or considerations. Usage Note: Some statistics will be relevant for numeric data, but not for strings. | | |
| | Statistic | User | Date | Stars | Title | Text | | --- | --- | --- | --- | --- | --- | | count | 315 | 315 | 308 | 315 | 315 | | mean |  |  | 8 |  | 92 parole | | min |  |  | 0 |  |  | | max |  |  | 10 |  |  |   **Above:** Queste sono le statistiche descrittive; purtroppo la maggior parte dei dati sono di tipo Text | | |

| Example of Data Points | | |
| --- | --- | --- |
| PRIMARY DATA MODALITY | SAMPLING OF DATA POINTS | DATA FIELDS |
| Text Data | No sampling (full data) | | **Field Name** | **Field Value** | **Description** | | --- | --- | --- | | User | Text | User name | | Stars | Text | Numero da 1-10 per classificare | | Text | Text | recensione | | Title | Text | Titolo della recensione | | Date | date | Data in cui è stata scritta la recensione | |
| TYPICAL DATA POINT | | A TYPICAL DATA POINT |
| The text on the right represents a single-line excerpt | | User: Britsahil123Stars: 10Text: I couldn't stop binging this! It's a funny and wild ride with an ensemble cast…. Title: Funny, brilliant and wild  Date: 7 March 2024 |

###### 

| Provenance | | |
| --- | --- | --- |
| Collection | | |
| METHOD(S) USED | METHODOLOGY DETAIL(S) | SOURCE DESCRIPTION(S) |
| Select **all applicable** methods used to collect data: | Provide a description of each collection method used.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for collection method type.)* | Provide a description of each upstream source of data.Use additional notes to capture any other relevant information or considerations. |
| Scraping del web | **Source:**https://www.imdb.com/title/tt13210838/reviews/?ref\_=tt\_ov\_rt  **Platform:** IMDb  **Is this source considered sensitive or high-risk?** No  **Dates of Collection:** [Maggio 2024]  **Primary modality of collected data:**  Text Data  **Update Frequency for collected data:**  static |  |

###### 

| Motivations & Intentions | | |
| --- | --- | --- |
| Motivations | | |
| PURPOSE(S) | DOMAIN(S) OF APPLICATION | MOTIVATING FACTOR(S) |
| TrainingTestingValidation | Sentiment analysis | *- Analizzare il sentiment della serie tv per capire se questo tipo di format può piacere al pubblico* |
| Intended Use | | |
| DATASET USE(S) | SUITABLE USE CASE(S) | UNSUITABLE USE CASE(S) |
| Select **one**: | Summarize known suitable and intended use cases of this dataset.Use additional notes to capture any specific patterns that readers should look out for, or other relevant information or considerations. | Summarize known unsuitable and unintended use cases of this dataset.Use additional notes to capture any specific patterns that readers should look out for, or other relevant information or considerations. |
| Safe for research use | **[Ricerca di mercato]** : Utilizzando dati demografici o di mercato, si possono condurre ricerche di mercato per comprendere meglio i bisogni dei consumatori o identificare nuove opportunità di business.  **[Valutazione della soddisfazione del cliente]**: si puo utilizzare per valutare la soddisfazione del cliente e identificare aree di miglioramento dei prodotti o servizi. | **[violazione della privacy ]** : il dataset include informazioni personali o sensibili, potrebbe essere vietato l'uso che potrebbe compromettere la privacy degli individui.  **[Violazione delle normative]**: Se l'uso del dataset viola le normative locali, nazionali o internazionali, potrebbe essere considerato inaccettabile. |
|  |  |  |

| Access, Retention, & Wipeout | | |
| --- | --- | --- |
| Access | | |
| ACCESS TYPE | DOCUMENTATION LINK(S) | PREREQUISITE(S) |
| Select **one**: | Provide links that describe documentation to access this dataset: | Please describe any required training or prerequisites to access this dataset. |
| External - Open Access | https://www.imdb.com/title/tt13210838/reviews/?ref\_=tt\_ov\_rt | *None* |

|  | Provenance | | |
| --- | --- | --- | --- |
|  | Collection | | |
|  | METHOD(S) USED | METHODOLOGY DETAIL(S) | SOURCE DESCRIPTION(S) |
|  | Select **all applicable** methods used to collect data: | Provide a description of each collection method used.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for collection method type.)* | Provide a description of each upstream source of data.Use additional notes to capture any other relevant information or considerations. |
|  | Scraped | **<scraping>** **Source:** <https://www.imdb.com/title/tt13210838/reviews/?ref\_=tt\_ov\_rt.>  **Platform:** [IMDb], <piattaforma che racchiude informazioni e recensioni riguardo I film e serie tv.>  **Is this source considered sensitive or high-risk?** [ No]  **Dates of Collection:** [giugno 2024-giugno 2024]  **Primary modality of collected data:**  *Usage Note: Select one for this collection type.*  Text Data  **Update Frequency for collected data:** *Usage Note: Select one for this collection type.*  static |  |
| COLLECTION CADENCE | | DATA INTEGRATION | DATA PROCESSING |
| Select **all applicable**: | | List all fields collected from different sources, and specify if they were included or excluded from the dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each upstream source.)* | Summarize how data from different sources or methods aggregated, processed, or connected.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each source OR collection method.)* |
|  | Static (Data was collected once from single or multiple sources.) | **<Source>**  **Included Fields**  (Data fields that were collected and are included in the dataset.)   | **Field Name** | **Description** | | --- | --- | | <Stars> | <numero di stelle assegnato, es: 8 ssu 10.> | | <Text> | <recensioni scritte dagli utenti.> | | … | … | | **<Collection Method or Source>**  **Description:** <si trovano tutte sullo stesso link ma bisogna pigiare il bottone LoadMore per caricare altre recensioni.>  **Methods employed:** <Scraping.>  **Tools or libraries:** <Selenium e Beautiful Soup.> |
|  | Collection Criteria | | |
|  | DATA SELECTION | DATA INCLUSION | DATA EXCLUSION |
|  | Summarize the data selection criteria.Use additional notes to capture any other relevant information or considerations. | Summarize the data inclusion criteria.Use additional notes to capture any other relevant information or considerations. | Summarize the data exclusion criteria.Use additional notes to capture any other relevant information or considerations. |
|  | **[Collection Method or Source]:** <ho utilizzato lo scraping del web scegliendo cosa mi sarebbe servito per fare la sentiment analysis.> | **[Collection Method or Source]:** < ho utilizzato lo scraping del web scegliendo cosa mi sarebbe servito per fare la sentiment analysis> | **none** |
|  | Relationship to Source | | |
|  | USE & UTILITY(IES) | BENEFIT AND VALUE(S) | LIMITATION(S) AND TRADE-OFF(S) |
|  | Describe how the resulting dataset is aligned with the purposes, motivations, or intended use of the upstream source(s).Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each source type.)* | Summarize the benefits of the resulting dataset to its consumers, compared to the upstream source(s).Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each source type.)* | What are the limitations of the resulting dataset to its consumers, compared to the upstream source(s)?Break down by source type. *(Usage Note: Duplicate and complete the following for each source type.)* |
|  | **Recensioni su IMDb**: Il dataset risultante include recensioni raccolte tramite scraping da IMDb. Le recensioni sono state selezionate in base ai criteri pertinenti per l'analisi del sentiment, ad esempio la presenza di un numero significativo di recensioni e una distribuzione rappresentativa delle valutazioni | **Recensioni su IMDb:** Il dataset risultante offre diversi benefici rispetto alle recensioni grezze su IMDb. Prima di tutto, le recensioni sono state preprocessate e pulite per rimuovere informazioni non rilevanti e standardizzarle per l'analisi del sentiment. Inoltre, il dataset potrebbe includere etichette aggiuntive per il sentiment derivante da algoritmi di analisi del linguaggio naturale, rendendo più facile per i consumatori utilizzare le recensioni per scopi specifici come la valutazione della soddisfazione dei clienti o l'analisi del feedback del prodotto | **Recensioni su IMDb**: Nonostante il dataset risultante offra numerosi benefici, potrebbe anche presentare alcune limitazioni rispetto alle recensioni grezze su IMDb. Ad esempio, durante il processo di scraping e preprocessing delle recensioni, potrebbero essere state perse alcune informazioni importanti o dettagliate presenti nelle recensioni originali su IMDb. Inoltre, l'aggiunta di etichette di sentiment derivanti da algoritmi di analisi del linguaggio naturale potrebbe non essere completamente accurata o rappresentativa delle opinioni degli utenti come le recensioni originali su IMDb. |
|  | Version and Maintenance | | |
|  | Fill this next row if: this is not the first version of the dataset, and there is no data card available for the first version. | | |
|  |  | FIRST VERSION | NOTE(S) AND CAVEAT(S) |
|  |  | Provide a **basic description of the first version** of this dataset. | Summarize the caveats or nuances of the first version of this dataset that may affect the use of the current version.Use additional notes to capture any other relevant information or considerations. |
|  |  | **Release date:** Maggio/2024  **Link to dataset:** [RECENSIONI.csv + 1 (versione)]  **Status:** [Select one: Limited Maintenance ]  **Size of Dataset:** 307 KB  **Number of Instances:** 315 | La colonna delle stelle contine solo 308 osservazioni, percò le righe senza stelle verranno eliminate in quanto le stelle servono per la creazione della soglia. |
|  | CADENCE | LAST AND NEXT UPDATE(S) | CHANGES ON UPDATE(S) |
|  | Select **one**. | Please describe the update schedule | Summarize the changes that occur when the dataset is refreshed.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each source type.)* |
|  | Static | **Date of last update:** 20/05/2024  **Total Data points affected:** 315  **Data points updated:** 315  **Data points added:**315  **Data points removed:** 0 | **Recensioni su IMDb**: Quando il dataset viene aggiornato, vengono rilevate e integrate nuove recensioni da IMDb. Le recensioni esistenti potrebbero essere aggiornate o corrette, ad esempio se vengono segnalati errori o se vengono apportate modifiche alla fonte originale. Le modifiche alle recensioni possono includere aggiunte, modifiche di testo o rimozioni in base alla loro disponibilità o alla qualità delle fonti. Inoltre, potrebbero essere state apportate modifiche ai metadati o ai campi aggiuntivi nel dataset per migliorare la qualità e la completezza delle informazioni. |

| Extended Use | | |
| --- | --- | --- |
| Use with Other Data | | |
| SAFETY LEVEL | KNOWN SAFE DATASET(S) OR DATA TYPE(S) | BEST PRACTICES |
| Select **one**: | List the known datasets or data types and corresponding transformations that **are safe to join or aggregate** this dataset with. | Summarize best practices for using this dataset with other datasets or data types.Use additional notes to capture any other relevant information or considerations. |
| Unknown | **Dataset demografico dei clienti:** Un dataset demografico dei clienti contiene informazioni sui clienti come età, genere e posizione geografica.  È possibile aggregare questo dataset con il dataset delle recensioni utilizzando l'identificatore univoco dell'utente per comprendere meglio i modelli di comportamento degli utenti in base alle loro caratteristiche demografiche. | Eseguire un'analisi preliminare dei dati per comprendere la loro struttura, distribuzione e qualità prima di combinarli con altri dataset. Questo aiuta a identificare eventuali problemi o anomalie nei dati e adottare misure correttive appropriate. |
|  | KNOWN UNSAFE DATASET(S) OR DATA TYPE(S) | LIMITATION(S) AND RECOMMENDATION(S) |
| Fill out this row if you selected “Conditionally safe to use with other datasets” or “Should not be used with other datasets”: | List the known datasets or data types and corresponding transformations that are **unsafe to join or aggregate** with this dataset. | Summarize limitations of the dataset that introduce foreseeable risks when the dataset is conjoined with other datasets.Use additional notes to capture any other relevant information or considerations. |
|  | **None** | **Dati sensibili o personali**: Se il dataset contiene informazioni sensibili o personali come dati finanziari, informazioni sanitarie o dettagli di identificazione personale nelle recensioni o titolo , non è sicuro unire o aggregare questi dati con altri dataset senza le adeguate precauzioni per garantire la privacy e la sicurezza dei dati. |
| Forking & Sampling | | |
| SAFETY LEVEL | ACCEPTABLE SAMPLING METHOD(S) | BEST PRACTICE(S) |
| Select **one**: | Select **all applicable** acceptable methods to sample this dataset: | Summarize the best practices for forking or sampling this dataset.Use additional notes to capture any other relevant information or considerations. |
| Safe to form and/or sample | Cluster Sampling  Weighted Sampling  Unsampled | **Validare i risultati**: Validare i risultati dell'analisi eseguita sui dati forked o campionati per assicurarsi che siano coerenti con le aspettative e che riflettano accuratamente i modelli presenti nel dataset originale. Utilizzare tecniche di convalida incrociata o confrontare i risultati con il dataset completo per confermare la validità delle conclusioni |
|  | RISK(S) AND MITIGATION(S) | LIMITATION(S) AND RECOMMENDATION(S) |
| Fill out this row if you selected “Conditionally safe to fork and/or sample” or “Should not be forked and/or sampled”. | Summarize known or residual risks associated with forking and sampling methods when applied to the dataset.Use additional notes to capture any other relevant information or considerations. | Summarize the limitations that the dataset introduces when forking or sampling the dataset and corresponding recommendations.Use additional notes to capture any other relevant information or considerations. |
|  | **Rischio di perdita di rappresentatività**: Esiste il rischio che la forcellazione o il campionamento del dataset possano portare alla perdita di rappresentatività dei dati, specialmente se i criteri di selezione non sono appropriati o se i campioni estratti non riflettono accuratamente la diversità del dataset originale. **Mitigation**: Per mitigare questo rischio, è importante utilizzare criteri di selezione appropriati che riflettano la struttura e la complessità del dataset originale. Inoltre, è consigliabile eseguire un'analisi della distribuzione delle variabili chiave nei dati selezionati per assicurarsi che siano rappresentativi | **Limitazione sulla rappresentatività dei campioni**: Esiste il rischio che i campioni estratti durante la forcellazione o il campionamento potrebbero non essere completamente rappresentativi del dataset originale, specialmente se i criteri di selezione non sono accurati o se il campionamento non è casuale. **Recommendation**: Utilizzare criteri di selezione accurati e obiettivi per garantire che i campioni estratti siano rappresentativi del dataset originale. Inoltre, quando possibile, utilizzare tecniche di campionamento casuale per ridurre il rischio di bias nei dati selezionati. |
| Use in ML or AI Systems | | |
| DATASET USE(S) | NOTABLE FEATURE(S) | USAGE GUIDELINE(S) |
| Select **all applicable**: | Describe any notable feature distributions or relationships between individual instances made explicit.Include links to servers where readers can explore the data on their own. | Summarize usage guidelines or policies that consumers should be aware of.Use additional notes to capture any other relevant information or considerations. |
| TrainingTestingValidationVettorizzazione Definizione soglia sentiment | **Stars (Stelle)**: La distribuzione delle stelle assegnate dalle recensioni potrebbe fornire insight sulla soddisfazione degli utenti con il prodotto o il servizio. Ad esempio, potrebbe essere interessante esaminare la distribuzione delle stelle per identificare se la maggior parte delle recensioni è positiva, neutra o negativa  **Additional Notes:**La visualizzazione della distribuzione delle stelle potrebbe aiutare a identificare tendenze o pattern significativi nelle recensioni | **Riservatezza e Sicurezza dei Dati**: Gli utenti sono responsabili della protezione e della sicurezza dei dati durante l'accesso, l'uso e la condivisione del dataset. È importante adottare misure di sicurezza adeguate per proteggere i dati sensibili o personali e rispettare le normative sulla privacy applicabili  **Approvazione dei Dati**: Prima di utilizzare o condividere il dataset, potrebbe essere necessario ottenere l'approvazione o l'autorizzazione da parte dell'ente o dell'organizzazione proprietaria del dataset. Ciò potrebbe implicare la compilazione di moduli di richiesta di accesso ai dati o il completamento di procedure di approvazione specifiche.  **Revisore dei Dati**: Valentina Baldi è responsabile della revisione e della supervisione dell'uso e della condivisione del dataset. Gli utenti possono fare riferimento a [Nome del revisore] per domande o chiarimenti relativi alle politiche di utilizzo dei dati o per richiedere assistenza nell'interpretazione dei dati.  **Additional Notes:** <Add here> |
|  | DISTRIBUTION(S) | KNOWN CORRELATION(S) |
|  | Describe the recommended splits and corresponding criteria.Use additional notes to capture any other relevant information or considerations. | Summarize any known correlations with the indicated features in this dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate for each known correlation.)* |
|  | Train 80%  Test 20% | Stars,Text  **Description:** La correlazione tra il numero di stelle assegnate e il testo delle recensioni potrebbe indicare se le recensioni con valutazioni più alte contengono linguaggio più positivo o se le recensioni negative hanno tendenzialmente un linguaggio più critico.  **Impact on dataset use**: Questa correlazione potrebbe influenzare l'analisi delle recensioni, consentendo agli utenti di identificare i principali temi o sentimenti associati alle valutazioni positive o negative**.**  **Risks from correlation:** Il rischio principale è che la correlazione tra il numero di stelle e il testo delle recensioni potrebbe portare a una sovrapposizione di informazioni, dove le stesse informazioni sono già implicitamente presenti nelle valutazioni stellari. Ciò potrebbe influenzare l'interpretazione dei risultati dell'analisi delle recensioni |
| SPLIT STATISTICS | | |
| Provide the sizes of each split. As appropriate, provide any descriptive statistics for features. | | |
| | Statistic | Train | Test | | --- | --- | --- | | Count | 246 | 62 | | positive | 167 | 40 | | negative | 79 | 22 |   **Above:** general info about train and test | | |

| Transformations | | |
| --- | --- | --- |
| Ho aggiunto la colonna Sentiment , calcolando la soglia del sentiment attraverso la colonna Stars | | |
| Synopsis | | |
| TRANSFORMATION(S) APPLIED | FIELD(S) TRANSFORMED | LIBRARY(IES) AND METHOD(S) USED |
| Select **all applicable** transformations that were applied to the dataset. | Provide the fields in the dataset that were transformed.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each transformation type applied. Include the data types to which fields were transformed.)* | Provide a description of the methods used to transform or process the dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each transformation type applied.)* |
| Cleaning Missing ValuesCreazione colonna tramite altra/e colonna/e | **<Transformation Type>**   | **Field Name** | **Source & Target** | | --- | --- | | <Stars> | Eliminazione righe senza parametro stars | | <Sentiment> | Creata utilizzando la colonna Stars | | … | … | | **Eliminazione righe senza parametro stars**  **Method:** Le righe con valori mancanti nella colonna "Stars" sono state eliminate per assicurare che solo le recensioni con valutazioni numeriche siano incluse nell'analisi.  **Platforms, tools, or libraries:**Python pandas library: Utilizzato per filtrare le righe con valori mancanti nella colonna "Stars".  L'eliminazione delle righe senza il parametro "Stars" ha garantito che solo le recensioni con valutazioni numeriche siano considerate nell'analisi del sentiment.  **Creazione del campo Sentiment utilizzando la colonna Stars**  **Method:** Il campo "Sentiment" è stato creato utilizzando la colonna "Stars", dove le recensioni con valutazioni di 4 o più stelle sono state etichettate come "positivo", mentre le recensioni con valutazioni inferiori a 4 stelle sono state etichettate come "negativo".  **Platforms, tools, or libraries:**Python pandas library: Utilizzato per creare la colonna "Sentiment" in base ai valori della colonna "Stars". |
| Breakdown of Transformations | | |
| Fill out relevant rows. | | |
| CLEANING MISSING VALUE(S) | METHOD(S) USED | COMPARATIVE SUMMARY |
| Which fields in the data were missing values? How many? | How were missing values cleaned? What other choices were considered? | Why were missing values cleaned using this method (over others)? Provide comparative charts showing before and after missing values were cleaned. |
| **Stars:** 7 valori non erano stati inseriti dagli user | Ho eliminato le righe senza stars  **Platforms, tools, or libraries:** Python pandas library: Utilizzato per filtrare le righe con valori mancanti nella colonna "Stars". | Perchè altrimenti non si poteva svolgere la sentiment analysis perchè non si sarebbero potute creare le righe della colonna sentiment per quei dati. |

| Annotations & Labeling | | |
| --- | --- | --- |
| Fill this section if any human or algorithmic annotation tasks were performed in the creation of your dataset. | | |
| ANNOTATION WORKFORCE TYPE | ANNOTATION CHARACTERISTIC(S) | ANNOTATION DESCRIPTION(S) |
| Select **all applicable** annotation workforce types or methods used to annotate the dataset: | Describe relevant characteristics of annotations as indicated. For quality metrics, consider including accuracy, consensus accuracy, IRR, XRR at the appropriate granularity (e.g. across dataset, by annotator, by annotation, etc.).Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each annotation type.)* | Provide descriptions of the annotations applied to the dataset. Include links and indicate platforms, tools or libraries used wherever possible.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each annotation type.)* |
| Machine-Generated (web scraping) | **(Annotation Type)**  Number of unique annotations: 1 (poiché le annotazioni sono generate automaticamente dal processo di web scraping  Total number of annotations : è uguale al numero di dati raccolti tramite web scraping  Avg. Annotations per example: Poiché le annotazioni sono generate automaticamente, non c'è un numero medio di annotazioni per esempio da considerare  Number of annotators per example: Non applicabile (poiché le annotazioni sono generate automaticamente, non coinvolgono annotatori umani)  [Quality metric per granularity]: Non applicabile (poiché le annotazioni sono generate automaticamente, non si applicano metriche di qualità basate su annotatori umani) | **Description:** Le annotazioni in questo dataset sono generate automaticamente attraverso il processo di web scraping. Gli script di scraping vengono utilizzati per estrarre dati specifici dai siti web di origine, come recensioni, valutazioni o altre informazioni rilevanti, e quindi strutturare queste informazioni in un formato utilizzabile nel dataset.  **Platforms, tools, or libraries:**  Python BeautifulSoup: Utilizzato per l'estrazione dei dati dai siti web di origine.  Python requests: Utilizzato per inviare richieste HTTP ai server dei siti web di origine e ottenere le pagine web da analizzare.  Python pandas: Utilizzato per la manipolazione e la strutturazione dei dati estratti in un formato tabellare utilizzabile. |
|  | ANNOTATION DISTRIBUTION(S) | ANNOTATION TASK(S) |
|  | Provide a distribution of annotations for each annotation or class of annotations using the format below.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each annotation type.)* | Summarize each task type associated with annotations in the dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each task type.)* |
|  | **(Annotation Type)**  Positive (1) 207  Negative (0) 101 | **Sentiment Analysis**  **Task description:** L'obiettivo di questo compito è classificare la polarità delle recensioni in positive o negative  **Task instructions:** Gli annotatori sono addestrati a leggere attentamente le recensioni e determinare se esprimono una valutazione positiva o negativa del prodotto o del servizio.  **Methods Used:** L'annotazione del sentiment potrebbe essere stata effettuata manualmente da annotatori umani che leggono e valutano il contenuto delle recensioni. Potrebbe anche essere stata utilizzata l'analisi automatica del sentiment utilizzando algoritmi di machine learning per classificare automaticamente le recensioni in positive o negative. |

| Validation Types | | |
| --- | --- | --- |
| Fill this section if the data in the dataset was validated during or after the creation of your dataset. | | |
| METHOD(S) | BREAKDOWN(S) | DESCRIPTION(S) |
| Select **all applicable**: | Provide a description of the fields and data points that were validated.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each validator type.)* | Provide a description of the methods used to validate the dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each validator type.)* |
| Cross- Validation stratificata |  | **(Cross-validation stratificata)**  **Method:** tecnica di validazione dei modelli di machine learning che garantisce una distribuzione uniforme delle classi target nei fold di addestramento e test. Ciò aiuta a migliorare l'affidabilità delle stime delle prestazioni del modello, specialmente quando si lavora con dataset sbilanciati.  **Platforms, tools, or libraries:**  **scikit-learn (sklearn)**: fornisce implementazioni efficienti di molte tecniche di cross validation, inclusa la StratifiedKFold per la cross validation stratificata  **numpy (np)**: Utilizzata per operazioni matematiche efficienti e manipolazione dei dati, spesso utilizzata in combinazione con scikit-learn per la gestione dei dati durante la cross validation |

###### 

###### 

###### 

| ML APPLICATION(S) | EVALUATION PROCESS(ES) |
| --- | --- |
| Provide a list of key ML tasks that the dataset has been used for. Usage Note: Use comma-separated keywords. | Provide a description of the evaluation process for the model's overall performance or the determination of how the dataset contributes to the model's performance.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each model and method used.)* |
| *Classification, Regression, Natural Language Processing (NLP), Text Classification, Sentiment Analysis, Deep Learning, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), BERT, GPT-2.*  *Classification, Regression, Natural Language Processing (NLP), Text Classification, Sentiment Analysis, Deep Learning, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), BERT, GPT-2.* | **NO SMOTE:**  **Vettorizzazione: BoW**  Modello: Logistic Regression  Accuracy (mean): 0.8171428571428571  Accuracy (std): 0.031127634228469567  Precision (mean): 0.8188104007255854  Precision (std): 0.04673626379943468  Recall (mean): 0.9467023172905525  Recall (std): 0.05395625053969607  F1-score (mean): 0.8756520632766444  F1-score (std): 0.01962450778131024  Modello: Naive Bayes  Accuracy (mean): 0.8087346938775509  Accuracy (std): 0.05313000252067309  Precision (mean): 0.8262324929971989  Precision (std): 0.05503712710193044  Recall (mean): 0.9158645276292334  Recall (std): 0.04037580712913113  F1-score (mean): 0.8672322166041294  F1-score (std): 0.03448124511757865  Modello: SVM  Accuracy (mean): 0.6748571428571428  Accuracy (std): 0.010847010770202826  Precision (mean): 0.6775510204081632  Precision (std): 0.008163265306122458  Recall (mean): 0.9941176470588236  Recall (std): 0.011764705882352944  F1-score (mean): 0.8058183955333529  F1-score (std): 0.007706488022641702  Modello: Random Forest  Accuracy (mean): 0.7155102040816327  Accuracy (std): 0.035899809677218  Precision (mean): 0.7099123053026707  Precision (std): 0.03375957636410399  Recall (mean): 0.9880570409982175  Recall (std): 0.01462979289032669  F1-score (mean): 0.8254563320681451  F1-score (std): 0.017953044056815262  Modello: KNN  Accuracy (mean): 0.7035102040816327  Accuracy (std): 0.0768256516927614  Precision (mean): 0.724031524031524  Precision (std): 0.046215369899143904  Recall (mean): 0.9098039215686274  Recall (std): 0.06607975611606508  F1-score (mean): 0.8060071850004225  F1-score (std): 0.05206519522889057  **Vettorizzazione: TF-IDF**  Modello: Logistic Regression  Accuracy (mean): 0.6788571428571428  Accuracy (std): 0.007924676636415747  Precision (mean): 0.6788571428571428  Precision (std): 0.007924676636415747  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8086870128598015  F1-score (std): 0.005592405117558078  Modello: Naive Bayes  Accuracy (mean): 0.6788571428571428  Accuracy (std): 0.007924676636415747  Precision (mean): 0.6788571428571428  Precision (std): 0.007924676636415747  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8086870128598015  F1-score (std): 0.005592405117558078  Modello: SVM  Accuracy (mean): 0.6788571428571428  Accuracy (std): 0.007924676636415747  Precision (mean): 0.6788571428571428  Precision (std): 0.007924676636415747  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8086870128598015  F1-score (std): 0.005592405117558078  Modello: Random Forest  Accuracy (mean): 0.7155102040816327  Accuracy (std): 0.030912810169406505  Precision (mean): 0.7053121532300863  Precision (std): 0.025689827598329124  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8269332463468502  F1-score (std): 0.01732213418030444  Modello: KNN  Accuracy (mean): 0.7602448979591836  Accuracy (std): 0.03442885571420076  Precision (mean): 0.7579731379731379  Precision (std): 0.024993080780991137  Recall (mean): 0.952584670231729  Recall (std): 0.039740452609921505  F1-score (mean): 0.8435738450446454  F1-score (std): 0.02202300566539481  **Vettorizzazione: Word2Vec**  Modello: Logistic Regression  Accuracy (mean): 0.6788571428571428  Accuracy (std): 0.007924676636415747  Precision (mean): 0.6788571428571428  Precision (std): 0.007924676636415747  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8086870128598015  F1-score (std): 0.005592405117558078  Modello: SVM  Accuracy (mean): 0.6788571428571428  Accuracy (std): 0.007924676636415747  Precision (mean): 0.6788571428571428  Precision (std): 0.007924676636415747  Recall (mean): 1.0  Recall (std): 0.0  F1-score (mean): 0.8086870128598015  F1-score (std): 0.005592405117558078  Modello: Random Forest  Accuracy (mean): 0.6868571428571429  Accuracy (std): 0.04086397219116512  Precision (mean): 0.7002415458937198  Precision (std): 0.013856189189527638  Recall (mean): 0.9397504456327986  Recall (std): 0.06356817149484024  F1-score (mean): 0.8020039922211657  F1-score (std): 0.03259158301328487  Modello: KNN  Accuracy (mean): 0.6342040816326531  Accuracy (std): 0.04595737118983798  Precision (mean): 0.7073146223146223  Precision (std): 0.032888878362327255  Recall (mean): 0.7903743315508022  Recall (std): 0.05079869894847937  F1-score (mean): 0.7455517130859596  F1-score (std): 0.03134507342268051  **SMOTE:**  **Vettorizzazione: BoW**  Modello: Logistic Regression  Accuracy (mean): 0.8413387607417457  Precision (mean): 0.8684650316930629  Recall (mean): 0.8085561497326204  F1-score (mean): 0.8334721495867704  Modello: Naive Bayes  Accuracy (mean): 0.8592944369063771  Precision (mean): 0.8437195484254307  Recall (mean): 0.886096256684492  F1-score (mean): 0.8633786520872828  Modello: SVM  Accuracy (mean): 0.8353686114880144  Precision (mean): 0.76590472549966  Recall (mean): 0.970409982174688  F1-score (mean): 0.8554789199852492  Modello: Random Forest  Accuracy (mean): 0.8056083220262324  Precision (mean): 0.8459518792191206  Recall (mean): 0.7484848484848484  F1-score (mean): 0.7914559094275119  Modello: KNN  Accuracy (mean): 0.5029850746268656  Precision (mean): 0.2  Recall (mean): 0.006060606060606061  F1-score (mean): 0.011764705882352943  **Vettorizzazione: TF-IDF**  Modello: Logistic Regression  Accuracy (mean): 0.9341022161917684  Precision (mean): 0.9132719430552247  Recall (mean): 0.9638146167557933  F1-score (mean): 0.9359687403833794  Modello: Naive Bayes  Accuracy (mean): 0.9251922207146087  Precision (mean): 0.9560483870967742  Recall (mean): 0.8921568627450981  F1-score (mean): 0.9226678697646438  Modello: SVM  Accuracy (mean): 0.8983717774762552  Precision (mean): 0.83335656213705  Recall (mean): 1.0  F1-score (mean): 0.9084138784093018  Modello: Random Forest  Accuracy (mean): 0.8715061058344642  Precision (mean): 0.8129755538579069  Recall (mean): 0.9759358288770054  F1-score (mean): 0.8852709558192668  Modello: KNN  Accuracy (mean): 0.5  Precision (mean): 0.0  Recall (mean): 0.0  F1-score (mean): 0.0  **Vettorizzazione: Word2Vec**  Modello: Logistic Regression  Accuracy (mean): 0.5419267299864314  Precision (mean): 0.7120397428404441  Recall (mean): 0.5376114081996435  F1-score (mean): 0.46955802495774446  Modello: SVM  Accuracy (mean): 0.5479421076436002  Precision (mean): 0.5805116959064327  Recall (mean): 0.3233511586452763  F1-score (mean): 0.4125565356231867  Modello: Random Forest  Accuracy (mean): 0.8262777023971053  Precision (mean): 0.8356345897941824  Recall (mean): 0.813903743315508  F1-score (mean): 0.8209154146230417  Modello: KNN  Accuracy (mean): 0.6407960199004974  Precision (mean): 0.819496336996337  Recall (mean): 0.3345811051693405  F1-score (mean): 0.4716811594202898  (no SMOTE)  **RNN Model** - Accuracy: 0.5967741935483871, F1-score: 0.7474747474747475  **CNN Model** - Accuracy: 0.6451612903225806, F1-score: 0.7441860465116279  **LSTM Model** - Accuracy: 0.6451612903225806, F1-score: 0.7843137254901961 **BERT**  Training Loss: 0.4782  Validation Loss: 0.5040  Validation Accuracy: 0.6452  **GPT-2**  precision recall f1-score support  0 1.00 1.00 1.00 62  accuracy 1.00 62  macro avg 1.00 1.00 1.00 62  weighted avg 1.00 1.00 1.00 62 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

###### 

|  | DESCRIPTION(S) AND STATISTIC(S) | EXPECTED PERFORMANCE AND  KNOWN CAVEATS |
| --- | --- | --- |
| Duplicate this row as necessary for each model type | Provide a description of the model(s) and task(s) that this dataset has been used in.Use additional notes to capture any other relevant information or considerations.(Usage Note: Duplicate and complete the following for each model.) | Provide a description of the expected performance and known caveats of the models for this dataset.Use additional notes to capture any other relevant information or considerations. *(Usage Note: Duplicate and complete the following for each model.)* |
|  | **Vettorizzazione: BoW**  **Modello: Logistic Regression**  **Accuracy (mean): 0.8171428571428571**  **Accuracy (std): 0.031127634228469567**  **Precision (mean): 0.8188104007255854**  **Precision (std): 0.04673626379943468**  **Recall (mean): 0.9467023172905525**  **Recall (std): 0.05395625053969607**  **F1-score (mean): 0.8756520632766444**  **F1-score (std): 0.01962450778131024**  **Modello: Naive Bayes**  **Accuracy (mean): 0.8087346938775509**  **Accuracy (std): 0.05313000252067309**  **Precision (mean): 0.8262324929971989**  **Precision (std): 0.05503712710193044**  **Recall (mean): 0.9158645276292334**  **Recall (std): 0.04037580712913113**  **F1-score (mean): 0.8672322166041294**  **F1-score (std): 0.03448124511757865**  **Modello: SVM**  **Accuracy (mean): 0.6748571428571428**  **Accuracy (std): 0.010847010770202826**  **Precision (mean): 0.6775510204081632**  **Precision (std): 0.008163265306122458**  **Recall (mean): 0.9941176470588236**  **Recall (std): 0.011764705882352944**  **F1-score (mean): 0.8058183955333529**  **F1-score (std): 0.007706488022641702**  **Modello: Random Forest**  **Accuracy (mean): 0.7155102040816327**  **Accuracy (std): 0.035899809677218**  **Precision (mean): 0.7099123053026707**  **Precision (std): 0.03375957636410399**  **Recall (mean): 0.9880570409982175**  **Recall (std): 0.01462979289032669**  **F1-score (mean): 0.8254563320681451**  **F1-score (std): 0.017953044056815262**  **Modello: KNN**  **Accuracy (mean): 0.7035102040816327**  **Accuracy (std): 0.0768256516927614**  **Precision (mean): 0.724031524031524**  **Precision (std): 0.046215369899143904**  **Recall (mean): 0.9098039215686274**  **Recall (std): 0.06607975611606508**  **F1-score (mean): 0.8060071850004225**  **F1-score (std): 0.05206519522889057** | **(Model Name)**  **Expected Performance:** <Summarize here. Include links where available.>  **Known Caveats:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.>  **Additional Notes:** <Add here> |

###### 