GXG at EVALITA 2018: Overview of the Cross-Genre Gender Prediction in Italian

Agnese Camici

Università di Pisa, Pisa, Italia

a.camici1@studenti.unipi.it

1. Abstract

**English.** Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ac erat dapibus, eleifend purus at, posuere nisi. Phasellus molestie elementum laoreet. Suspendisse maximus tortor ac mi egestas eleifend. Aliquam faucibus dui nisi, nec porttitor lorem sagittis ultricies. Pellentesque varius lacinia dui eget bibendum. Vivamus id orci gravida, fringilla quam eu, lacinia diam. Quisque vulputate urna at suscipit bibendum. Donec eu neque magna. Curabitur eget felis tellus. Curabitur vel nisl et augue mollis rhoncus.

**Italiano.** Maecenas finibus ante arcu, sit amet scelerisque lacus viverra id. Pellentesque maximus venenatis quam aliquet faucibus. Curabitur tristique consequat purus non-bibendum. Pellentesque venenatis at dui in tincidunt. In vulputate libero sem, et molestie urna ultrices sit amet. Etiam vel condimentum magna. In fringilla arcu lectus, at aliquam nulla vehicula eu.

# Introduction

The following instructions are directed to authors of papers submitted to CLiC-it 2023 or accepted for publication in its proceedings. All authors are required to adhere to these specifications. Authors are required to provide a Portable Document Format (PDF) version of their papers. The proceedings are designed for printing on A4 paper.

# Dataset

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors’ names and affiliations, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Section 2.6). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 3. **Do not number the pages**.

## Struttura del dataset(?)

Your PDF can be prepared using LaTeX with the CLiC-it 2023 style file (clic2023.sty, adapted from the official ACL 2014 style file) and the ACL bibliography style (acl.bst). You can alternatively use Microsoft Word to produce your PDF file. In this case, we strongly recommend the use of the Word template file (clic2023.odt) on the CLiC-it 2023 website.

## Organizzazione dei file(?)

For the production of the electronic manuscript you must use Adobe’s Portable Document Format (PDF).

## Preprocessing

Il primo passo del progetto ha previsto il download del dataset ufficiale dal sito di EVALITA 2018. I file sono stati salvati all’interno di una directory denominata *dataset\_originale*, collocata nella sottocartella *data/* della struttura di progetto.

Successivamente, mediante lo script contenuto nel notebook *extract\_text\_from\_doc.ipynb*, sono stati estratti i testi contenuti nei file originali in formato .txt (ad esempio *CH\_train.txt, CH\_test.txt*, ecc.). Ogni testo individuale è stato salvato come file .txt separato nella rispettiva sottocartella di genere (Children, Diary, Journalism, Twitter, YouTube). Per poter utilizzare il tool Profiling-UD, è stata creata una cartella sorella di *text\_from\_docs*, denominata *profiling\_input/*, contenente cinque sottocartelle, una per ciascun genere testuale (Children, Diary, Journalism, Twitter, YouTube). All’interno di ognuna di esse sono stati copiati sia i testi di training che quelli di test relativi al genere corrispondente, già convertiti in formato .txt. Profiling-UD ha elaborato correttamente i dati relativi ai generi Diary, journalism e Twitter. Tuttavia, durante l’elaborazione dei dati per i generi Children e YouTube, si sono manifestati degli errori non specificati (“An error occurred”). Dopo una prima analisi dei log, è emerso che uno dei file di test del genere Children — in particolare il file con ID 172 — causava un errore sistematico. Per risolvere il problema, questo file è stato eliminato dalla directory profiling\_input/children e dal file delle etichette gold corrispondente (test\_CH.gold). Un ulteriore tentativo di identificare altri file problematici è stato condotto mediante uno script di debug, che ha segnalato altri documenti potenzialmente corrotti o mal formattati. Tuttavia, anche dopo la loro rimozione, l’errore durante l'elaborazione del genere YouTube persisteva. Pertanto, si è deciso di escludere il genere YouTube dal progetto, concentrando l’analisi sui restanti quattro generi: Children, Diary, Journalism e Twitter. Nel caso del genere Twitter, il file di test fornito (*test\_TW.gold*) includeva le etichette gold solo per un sottoinsieme dei testi presenti nel corpus. Per garantire coerenza nella valutazione del modello, è stato necessario filtrare i dati di test e mantenere solo i testi per cui fosse disponibile una classificazione manuale. A tal fine, è stato effettuato un parsing del file *test\_twitter.gold*, da cui sono stati estratti gli identificativi testuali dei documenti annotati. Questi identificativi sono stati utilizzati per creare una sottocartella dedicata, *data/profiling\_input/twitter\_matching\_gold/,* contenente esclusivamente i file *.conllu* e le feature linguistiche (*linguistic\_profile.csv*) dei testi con etichetta gold. Queste operazioni sono state fatte nel notebook *NLM\_InGenre-TW.ipynb*

# Modelli e risultati

## SVM con ProfilingUD

Il modello SVM lineare, addestrato su 138 feature linguistiche non lessicali estratte da Profiling-UD, viene valutato tramite una 5-fold cross validation su ciascun genere.

**Children.** I dati utilizzati includono 200 documenti per il training e 199 per il test, con ciascun fold composto da 160 esempi di training e 40 di test. Come visibile nella *Tabella 1*, le performance ottenute risultano altamente instabili tra i fold, con accuracy variabili tra 0.425 e 0.625, mentre le baseline dummy oscillano tra 0.4 e 0.475. In alcuni casi, il modello supera la baseline dummy, ma senza garantire una generalizzazione robusta.

|  |  |  |
| --- | --- | --- |
| **SVM ProfilingUD Children** | | |
| *fold* | *accuracy* | *Dummy classifier* |
| 1 | 0.625 | 0.475 |
| 2 | 0.425 | 0.425 |
| 3 | 0.45 | 0.475 |
| 4 | 0.525 | 0.4 |
| 5 | 0.5 | 0.425 |

Tabella 1: Accuracy per fold dell’SVM rispetto alla baseline dummy su Children (5-fold cross validation)

Quando il modello viene poi valutato sul test set ufficiale (*Tabella 5*), ottiene un’accuracy di 0.5729 e un F1-score macro di 0.57, superando la baseline dummy (0.5025) ma restando su valori moderati. La confusion matrix (*Tabella 5*) evidenzia uno squilibrio tra le due classi: il modello classifica correttamente 69 testi scritti da autrici (classe F), mentre solo 45 testi scritti da autori (classe M) vengono etichettati correttamente. La restante parte degli esempi maschili (54) viene confusa con la classe femminile, indicando una tendenza del classificatore a sbilanciarsi verso la classe F. Infine, l’analisi dei pesi appresi dal classificatore SVM evidenzia che le feature più rilevanti (*Figura 1*) per la distinzione tra generi appartengono principalmente all’ambito verbale e morfosintattico. Tra le variabili con il peso maggiore figurano, ad esempio: *verbal\_root\_perc, verbs\_mood\_dist\_Imp, subordinate\_dist\_4, dep\_dist\_cop.*

Immagine che contiene testo, schermata, numero, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 1: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Children

Questo indica che, nel processo di apprendimento, il modello si affida soprattutto a come vengono usati i verbi — includendo la modalità, il tempo, la persona e il tipo di costruzione sintattica — per separare i due profili di genere. Si tratta di un risultato coerente con l’ipotesi secondo cui le strategie sintattiche verbali possono variare tra scrittori e scrittrici, anche in modo non consapevole. Tuttavia, l’efficacia di queste feature si dimostra limitata nel caso specifico del genere Children. Questo accade perché i testi all’interno di questo dominio presentano uno stile molto omogeneo, indipendentemente dall’autore. La conseguenza è che il modello fatica a tracciare un confine netto

tra le due classi e non riesce a sfruttare pienamente le differenze sintattiche individuate. Le feature verbali, seppur informative, non bastano da sole a garantire una separabilità efficace tra i due profili.

**Diary.** I dati utilizzati includono 200 documenti per il training e 74 per il test, con ciascun fold della 5 fold cross validation composto da 160 esempi di training e 40 di validazione. Come visibile nella *Tabella 2*, le performance del modello risultano moderatamente stabili tra i fold, con accuracy comprese tra 0.425 e 0.6, mentre le baseline dummy oscillano tra 0.375 e 0.5. In alcuni casi, l’SVM supera nettamente la baseline, suggerendo una parziale capacità di generalizzazione, anche se non costante.

|  |  |  |
| --- | --- | --- |
| **SVM ProfilingUD Diary** | | |
| *fold* | *accuracy* | *Dummy classifier* |
| 1 | 0.575 | 0.425 |
| 2 | 0.6 | 0.475 |
| 3 | 0.5 | 0.375 |
| 4 | 0.575 | 0.5 |
| 5 | 0.425 | 0.475 |

Tabella 2: Accuracy per fold dell’SVM rispetto alla baseline dummy su Diary (5-fold cross validation)

Quando il modello viene valutato sul test set (*Tabella 5)* ottiene un’accuracy di 0.5541 e un F1-score macro di 0.55, superando la baseline (0.5) e mantenendo un livello di performance costante con la cross-validation. La confusion matrix (*Tabella 5*) mostra che il modello classifica correttamente 24 testi scritti da autori (classe M), contro 17 testi correttamente classificati come F. La classe femminile, con recall = 0.46 e F1 = 0.51, risulta essere quella più problematica, mentre la classe maschile è riconosciuta con maggiore successo (recall = 0.65, F1 = 0.59). Questo comportamento suggerisce una leggera preferenza del modello per la classe M, al contrario di quanto osservato nel dominio Children. L’analisi delle feature più rilevanti apprese dal classificatore (Figura Z) rivela che, a differenza del caso Children, il modello si basa in misura significativa su caratteristiche sintattiche meno centrali rispetto al verbo. Tra le feature con peso maggiore troviamo: *char\_per\_tok, dep\_dist\_vocative, prep\_dist\_5* e *dep\_dist\_expl\_impers*-

Immagine che contiene testo, schermata, numero, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 2: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Diary

Questo indica che, nel processo di apprendimento, il modello si affida a diverse proprietà strutturali del testo, come la densità morfosintattica, le connessioni tra elementi e la distribuzione dei ruoli sintattici, piuttosto che alle sole strutture verbali. Si tratta di un comportamento coerente con la natura del genere Diary, che tende a includere variazioni stilistiche più libere, con maggiore uso di vocativi, costruzioni dislocate e tratti soggettivi. Tuttavia, l’efficacia di queste feature resta contenuta: il modello non riesce ancora a separare con sicurezza i due profili.

**Journalism.** I dati utilizzati includono 200 documenti per il training e 200 per il test, con ciascun fold composto da 160 esempi di training e 40 di validazione. Come riportato nella *Tabella 3*, le performance del modello SVM risultano relativamente stabili e superiori alle baseline, con accuracy che variano da 0.55 a 0.70, mentre le baseline dummy oscillano tra 0.425 e 0.475. In ogni fold il modello riesce a superare il classificatore casuale, indicando una buona capacità di generalizzazione sul dominio giornalistico.

|  |  |  |
| --- | --- | --- |
| **SVM ProfilingUD Journalism** | | |
| *fold* | *accuracy* | *Dummy classifier* |
| 1 | 0.675 | 0.450 |
| 2 | 0.700 | 0.425 |
| 3 | 0.675 | 0.475 |
| 4 | 0.550 | 0.425 |
| 5 | 0.675 | 0.425 |

Tabella 3: Accuracy per fold dell’SVM rispetto alla baseline dummy su Journalism (5-fold cross validation)

Quando valutato sul test set (*Tabella 5*) l’SVM ottiene un’accuracy di 0.57 e un F1-score macro di 0.57, con una leggera flessione rispetto alla validazione incrociata. La confusion matrix (*Tabella 5*) mostra che il modello classifica correttamente 63 testi femminili e 51 testi maschili, ma confonde 49 esempi M con la classe F, suggerendo una tendenza parziale al sovra-classificare come F. Nonostante ciò, l’equilibrio tra precision e recall resta accettabile, e la performance si mantiene superiore alla baseline dummy (0.50). L’analisi delle 15 feature più rilevanti per il classificatore SVM (*Figura 3*) mostra una combinazione di tratti morfosintattici e strutturali. Tra le feature più pesate figurano: *dep\_dist\_parataxis, n\_prepositional\_chains, dep\_dist\_iobj* e *verbs\_form\_dist\_Inf*.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 3: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Journalism

Questo suggerisce che, per il genere Journalism, il modello si affida soprattutto a strutture sintattiche complesse e informative, come paratassi, catene preposizionali e verbi in forma non finita. Rispetto ad altri generi testuali, qui le differenze stilistiche tra M e F si esprimono in modo più marcato sulle strutture frasali e sul livello di subordinazione, elementi che il classificatore riesce a sfruttare. Tuttavia, l’F1-score moderato (*Tabella 5*) indica che, pur in presenza di segnali utili, la separazione tra i due profili non è ancora pienamente affidabile, e potrebbero servire feature aggiuntive o modelli più sofisticati per migliorare la discriminazione.

**Twitter.** I dati utilizzati includono 6000 documenti per il training e 152 per il test. Questi sono in numero sensibilmente inferiore rispetto ai dati di training a causa del fatto che, per Twitter, solo una parte dei documenti del test set aveva un corrispondente nel file .gold. Ciascun fold della cross-validation è composto da 4800 esempi di training e 1200 di validazione. Come mostrato nella *Tabella 5*, le performance del modello SVM risultano relativamente stabili tra i fold, con accuracy comprese tra 0.611 e 0.649, mentre le baseline dummy oscillano tra 0.474 e 0.497. In ogni fold, il modello supera la baseline, confermando una buona capacità di catturare segnali discriminanti anche in un dominio caratterizzato da testi molto brevi e informali.

|  |  |  |
| --- | --- | --- |
| **SVM ProfilingUD Twitter** | | |
| *fold* | *accuracy* | *Dummy classifier* |
| 1 | 0.632 | 0.491 |
| 2 | 0.649 | 0.497 |
| 3 | 0.612 | 0.478 |
| 4 | 0.615 | 0.498 |
| 5 | 0.622 | 0.474 |

Tabella 4: Accuracy per fold dell'SVM rispetto alla baseline dummy su Twitter (5-fold cross validation)

Quando valutato sul test set ufficiale (*Tabella 5*), il modello mostra una flessione significativa delle performance: l’accuracy scende a 0.47 e l’F1-score macro a 0.45, avvicinandosi alla baseline dummy (0.43). La confusion matrix (*Tabella 5*) evidenzia una forte asimmetria: il modello riconosce correttamente 51 F su 65 (recall F = 0.78), ma solo 21 M su 87 (recall M = 0.24), confondendo 66 esempi maschili con la classe femminile. Questo comportamento indica una tendenza a sovra-classificare i testi come F quando si passa dal training set ampio a un test set più ridotto e probabilmente meno rappresentativo. L’analisi delle feature più rilevanti () mostra che il modello si affida a variabili legate sia alla struttura sintattica che alla densità lessicale e morfosintattica. Tra le feature principali compaiono: *dep\_dist\_flat:name, lexical\_density, char\_per\_tok,* e *dep\_dist\_aux*.

Immagine che contiene testo, schermata, Carattere, numero

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 4: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Twitter

Questi risultati indicano che, anche in presenza di testi brevi e molto eterogenei come quelli di Twitter, l’SVM è in grado di sfruttare differenze stilistiche, seppur meno stabili rispetto ad altri generi. Tuttavia, la netta diminuzione delle performance sul test set suggerisce che il modello soffre la variabilità tipica del linguaggio social e fatica a generalizzare su dati meno rappresentativi. L’asimmetria osservata nella confusion matrix sottolinea la difficoltà del classificatore nel riconoscere i testi maschili in questo dominio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SVM ProfilingUD** | | | | |
|  | *accuracy* | *F1-score (macro avg)* | *support* | *confusion matrix* |
| **CH** | 0.5729 | 0.57 | 199 | [[69, 31], [54, 45]] |
| **DI** | 0.5541 | 0.55 | 74 | [[17, 20], [13, 24]] |
| **JO** | 0.57 | 0.57 | 200 | [[63, 37], [49, 51]] |
| **TW** | 0.47 | 0.45 | 152 | [[51, 14], [66, 21]] |

Tabella 5: Report finale sul test set di SVM ProfilingUD

**Valutazione complessiva dell’SVM con Profiling-UD**. Nel complesso, le performance del modello SVM basato su feature linguistiche non lessicali di Profiling-UD risultano piuttosto modeste e soggette a notevoli fluttuazioni a seconda del genere testuale. Sebbene il modello mostri una lieve superiorità rispetto alla baseline casuale su tutti i generi, i valori di accuracy e F1-score si mantengono costantemente bassi, raramente superando la soglia del 0.57, e in alcuni casi (come per Twitter) scendendo a livelli inferiori (F1 = 0.45, accuracy = 0.47). L’analisi delle confusion matrix conferma che la capacità di discriminazione tra i generi rimane limitata in ogni dominio: nei casi Journalism e Children, l’apparente bilanciamento tra le classi è dovuto più all’incapacità del modello di cogliere segnali realmente distintivi che a una vera robustezza predittiva. Nei generi Diary e soprattutto Twitter, emergono difficoltà ancora maggiori, con tendenze marcate alla classificazione errata (in particolare, una forte sovra-predizione della classe femminile su Twitter). Questi risultati suggeriscono che la strategia di affidarsi esclusivamente a feature morfosintattiche non lessicali – pur interessante dal punto di vista linguistico – non è sufficiente per una classificazione efficace del genere in testi di diversa tipologia. La scarsa generalizzazione, unita alla vulnerabilità a dati rumorosi o poco informativi (come i testi brevi di Twitter), evidenzia la fragilità del modello in assenza di segnali più marcati o di informazioni contestuali aggiuntive. Nel complesso, l’SVM con Profiling-UD appare adatto solo a una distinzione grossolana in domini relativamente strutturati, mentre risulta poco affidabile quando la variabilità stilistica o la brevità dei testi aumenta.

## SVM con word embeddings

In questa sezione è stato sviluppato un classificatore SVM lineare che prende in input una rappresentazione del testo costruita attraverso word embeddings. L’obiettivo è analizzare l’impatto di diverse strategie di rappresentazione dei testi. A tal fine sono state testate tutte le combinazioni tra:

* Due metodi di aggregazione: *mean* e *max*
* Tre filtri lessicali: *all, verb* e *noun\_adj*

Le sei strategie risultanti sono:

1. *mean\_all*
2. *mean\_verb*
3. *mean\_noun\_adj*
4. *max\_all*
5. *max\_verb, max\_noun\_adj.*

Ciascuna strategia è stata valutata tramite 5-fold cross-validation sul training set.

**Children.** Come visibile nella *Tabella 6,* la strategia *mean\_noun\_adj* si è distinta come la più efficace in validazione, con una accuracy media di 0.705, mostrando buona robustezza anche tra i fold. Al contrario, *max\_verb* si è rivelata la meno efficace, con performance instabili e inferiori alla media.

|  |  |
| --- | --- |
| **SVM word embeddings Children** | |
| *strategia* | *mean accuracy* |
| mean\_all | 0.675 |
| mean\_verb | 0.610 |
| **mean\_noun\_adj** | **0.705** |
| max\_verb | 0.545 |
| max\_all | 0.655 |
| max\_noun\_adj | 0.585 |

Tabella 6: Mean accuracy della cross validation di ciascuna strategia (genere Children)

Come da specifiche del progetto, sul test set è stata testata esclusivamente la strategia migliore, ovvero *mean\_noun\_adj*. I risultati presentati in *Tabella 7* mostrano un drastico calo rispetto alla media osservata durante la cross validation (0.705 🡪 0.3919), l’accuracy è addirittura inferiore alla baseline del dummy classifier (0.5025). Questo può essere attribuito a un fenomeno di overfitting del modello ai dati di training: infatti, la strategia ha mostrato una buona stabilità nella cross-validation, con accuracies pari a 0.6, 0.675, 0.675, 0.8 e 0.775.  È possibile che il classificatore abbia appreso caratteristiche specifiche dei fold di training, risultando poco generalizzabile su dati nuovi e leggermente differenti come quelli del test set.

La confusion matrix evidenzia una forte tendenza del modello a predire la classe maschile: ben 44 testi di genere femminile su 46 sono stati classificati erroneamente come maschili. Questa asimmetria nelle predizioni suggerisce una bassa capacità del classificatore di distinguere correttamente tra le due classi, a discapito della classe femminile. Una possibile causa aggiuntiva è legata alla natura stessa dei testi appartenenti al genere "Children", spesso brevi, semplici e con vocabolario limitato. Questo potrebbe ridurre l’efficacia degli embeddings mediati sui soli sostantivi e aggettivi, compromettendo la qualità della rappresentazione.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SVM word embeddings** | | | | |
|  | *strategia* | *accuracy* | *support* | *confusion matrix* |
| **CH** | *mean\_noun\_adj* | 0.3919 | 74 | [[2, 44], [1, 27]] |
| **DI** | *mean\_noun\_adj* | 0.5270 | 74 | [[36, 1], [34, 3]] |
| **JO** |  |  |  |  |
| **TW** |  |  |  |  |

Tabella 7: risultati dell'SVM con word embeddings sul test set delle migliori strategie per ciascun genere

## SVM con n-grammi

## NLM: BERT italiano

Infine, è stato sperimentato un approccio basato su modelli pre-addestrati di tipo transformer. In particolare, si è scelto di utilizzare BERT basato sul modello *bert-base-uncased*, scaricato e integrato tramite la libreria transformers di Hugging Face. Il dataset per ciascun genere è stato suddiviso in training e validation set, e i testi sono stati tokenizzati tramite *BertTokenizer* con le seguenti impostazioni:

* Max sequence length: 128 token
* Batch size: 16
* Numero di epoche: 5
* Ottimizzatore: *AdamW* (con weight decay)
* Learning rate: 2e-5

Durante il training sono stati monitorati i valori di loss di training, loss di validazione e macro F1-score a ogni epoca.

**Children**. L’andamento delle curve di loss durante le cinque epoche di addestramento evidenzia una progressiva diminuzione sia della training loss che della validation loss, come riportato in *Tabella 6* e visualizzato in

*Tabella* 1: Accuracy per fold dell’SVM rispetto alla baseline dummy su Children (5-fold cross validation) 2

Tabella 2: Accuracy per fold dell’SVM rispetto alla baseline dummy su Diary (5-fold cross validation) 3

Tabella 3: Accuracy per fold dell’SVM rispetto alla baseline dummy su Journalism (5-fold cross validation) 4

Tabella 4: Accuracy per fold dell'SVM rispetto alla baseline dummy su Twitter (5-fold cross validation) 4

Tabella 5: Report finale sul test set di SVM ProfilingUD 5

Tabella 6: Risultati del fine-tuning di BERT sul genere Children per 5 epoche 6

Tabella 7: Risultati del fine-tuning di BERT sul genere Diary per 5 epoche 7

Tabella 8: Risultati del fine-tuning di BERT sul genere journalism per 5 epoche 8

Tabella 9: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche 9

Tabella 10: Metriche di valutazione del modello BERT sui dati di test suddivisi per genere 9

.

|  |  |  |  |
| --- | --- | --- | --- |
| **BERT Children** | | | |
| *epoch* | *training loss* | *validation loss* | *F1* |
| 1 | 0.707600 | 0.701953 | 0.333333 |
| 2 | 0.672800 | 0.667023 | 0.648411 |
| 3 | 0.653400 | 0.659546 | 0.664732 |
| 4 | 0.588600 | 0.593158 | 0.750000 |
| 5 | 0.541200 | 0.575374 | 0.750000 |

Tabella 8: Risultati del fine-tuning di BERT sul genere Children per 5 epoche



Figura 5: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Children in 5 epoche

In particolare, si osserva una discesa regolare della validation loss da 0.7019 a 0.5753, con un andamento parallelo al training loss, che raggiunge un minimo di 0.5412 all’ultima epoca. Questo comportamento è generalmente indice di un processo di apprendimento stabile e di una buona generalizzazione, almeno nella fase di validazione. Il valore dell’F1 score (macro avg) durante la validazione cresce in modo significativo tra la prima e la seconda epoca (da 0.33 a 0.64), stabilizzandosi poi tra la terza e la quinta epoca a 0.75, valore che rappresenta il picco di performance raggiunto nel training. Tuttavia, nel momento in cui il modello viene valutato sul test set ufficiale, l’F1 macro scende a 0.63, mentre l’accuracy si attesta al 64%. Questo scostamento tra validazione e test suggerisce una parziale sovra-ottimizzazione sul validation set, ma allo stesso tempo conferma una buona capacità del modello di generalizzare oltre i dati di training. L’analisi della confusion matrix (*Figura 6*) evidenzia una migliore performance nella classificazione della classe F (74 esempi correttamente classificati su 100), rispetto alla classe M (53 su 99), che risulta più difficile da distinguere.

Immagine che contiene testo, schermata, diagramma, Rettangolo

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 6: Confusion matrix delle performance di BERT sul genere Diary

Questo sbilanciamento potrebbe riflettere delle caratteristiche stilistiche più marcate o riconoscibili nei testi femminili del genere Children, oppure una tendenza del modello a privilegiare la classe con pattern linguistici più stabili. Il valore dell’F1-score (macro avg) durante la validazione cresce in modo significativo tra la prima e la seconda epoca (da 0.33 a 0.64), stabilizzandosi poi tra la terza e la quinta epoca a 0.75, valore che rappresenta il picco di performance raggiunto nel training. Tuttavia, nel momento in cui il modello viene valutato sul test set ufficiale, l’F1 macro scende a 0.63, mentre l’accuracy si attesta al 64%. Questo scostamento tra validazione e test suggerisce una parziale sovra-ottimizzazione sul validation set, ma allo stesso tempo conferma una buona capacità del modello di generalizzare oltre i dati di training.

**Diary**. L’andamento delle curve di loss durante le cinque epoche di addestramento mostra un miglioramento progressivo e marcato delle performance, come evidenziato in *Tabella 7* e *Figura 7.*

|  |  |  |  |
| --- | --- | --- | --- |
| **BERT Diary** | | | |
| *epoch* | *training loss* | *validation loss* | *F1* |
| 1 | 0.668600 | 0.779911 | 0.171698 |
| 2 | 0.584300 | 0.592498 | 0.731111 |
| 3 | 0.333800 | 0.513956 | 0.778999 |
| 4 | 0195200 | 0.430207 | 0.826302 |
| 5 | 0.118700 | 0.461366 | 0.826302 |

Tabella 9: Risultati del fine-tuning di BERT sul genere Diary per 5 epoche

Immagine che contiene testo, linea, Diagramma, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 7: Curve di loss del training e della validation di BERT sul genere Diary nelle 5 epoche

La training loss cala in modo consistente da 0.6686 a 0.1187, mentre la validation loss registra una riduzione iniziale da 0.7799 a 0.4302 alla quarta epoca, seguita da un lieve peggioramento nell’ultima epoca (0.4614), che potrebbe indicare l'inizio di un leggero overfitting. Tuttavia, nel complesso, il comportamento del modello risulta stabile e ben generalizzato in fase di validazione. Il valore di F1-score (macro avg) aumenta significativamente già alla seconda epoca (0.73), per poi raggiungere un picco stabile di 0.8263 tra la quarta e la quinta epoca. Questo valore rappresenta la performance massima raggiunta dal modello sul validation set. Quando valutato sul test set ufficiale, BERT ottiene un'accuracy del 69% e un F1 macro di 0.68, mostrando un lieve calo rispetto alla validazione, ma mantenendo comunque un livello di performance elevato. L’analisi della confusion matrix (*Figura 8*) suggerisce una migliore classificazione della classe F (31 esempi corretti su 37, recall 0.84), rispetto alla classe M (20 su 37, recall 0.54), con una tendenza del modello a favorire la predizione della classe femminile.

Immagine che contiene testo, schermata, diagramma, Rettangolo

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 8: Confusion matrix delle performance di BERT sul genere Diary

Questo squilibrio potrebbe derivare da differenze stilistiche più marcate nei testi scritti da autrici, oppure da una distribuzione dei dati che rende la classe F più riconoscibile nel contesto del genere diary.

**Journalism**. Le curve di loss nel corso delle cinque epoche mostrano una riduzione progressiva del training loss, che scende da 0.6958 a 0.4013. La validation loss si stabilizza tra valori compresi tra 0.6636 e 0.679 dopo un iniziale calo nella seconda epoca. Come mostrato in *Tabella 8* e *Figura 9*, la validation loss non segue una traiettoria chiaramente decrescente, suggerendo una fase di apprendimento meno stabile rispetto agli altri generi.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| **BERT Journalism** | | | |
| *epoch* | *training loss* | *validation loss* | *F1* |
| 1 | 0.695800 | 0.731222 | 0.218182 |
| 2 | 0.619500 | 0.666098 | 0.600000 |
| 3 | 0.555300 | 0.676119 | 0.557460 |
| 4 | 0.465100 | 0.679120 | 0.617222 |
| 5 | 0.401300 | 0.663613 | 0.645614 |

Tabella 10: Risultati del fine-tuning di BERT sul genere journalism per 5 epoche

|  |
| --- |
| Figura 9: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Journalism in 5 epoche |

Tale andamento indica una possibile difficoltà del modello nel generalizzare su questo dominio testuale, con indizi iniziali di overfitting lieve a partire dalla terza epoca. L’F1-score in validazione cresce rapidamente tra la prima (0.22) e la seconda epoca (0.60), mantenendo valori intermedi nelle epoche successive, fino a un massimo di 0.65 alla quinta. Tuttavia, quando il modello viene valutato sul test set, le prestazioni calano sensibilmente: l’accuracy si attesta al 56%, e l’F1 macro scende a 0.56, valori che riflettono una generalizzazione debole e una capacità discriminativa inferiore rispetto agli altri generi. La confusion matrix (*Figura 10*) conferma questo andamento: su 100 testi per ciascuna classe, il modello classifica correttamente solo 52 testi della classe M e 61 della classe F, commettendo errori distribuiti quasi equamente.

Immagine che contiene testo, schermata, diagramma, Rettangolo

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 10: Confusion matrix delle performance di BERT sul genere journalism

Questa simmetria nell’errore suggerisce che BERT, nel dominio journalism, non riesce a identificare pattern linguistici sufficientemente forti per distinguere il genere dell’autore. Possibili motivazioni includono la maggiore neutralità stilistica tipica dei testi giornalistici, o una minore marcatura linguistica di genere in questo tipo di scrittura.

**Twitter.** Le curve di loss mostrano un comportamento indicativo di overfitting a partire dalla terza epoca: la training loss si riduce in modo continuo da 0.6272 a 0.1186, mentre la validation loss, dopo un iniziale miglioramento fino alla seconda epoca (0.5183), inizia a peggiorare rapidamente, raggiungendo valori molto alti (1.3494 alla quinta epoca). Come mostrato in *Tabella 9* e *Figura 11*, questo disallineamento crescente tra le curve suggerisce che il modello abbia appreso in modo eccessivo le peculiarità del training set, perdendo capacità di generalizzazione sui dati non visti.

|  |  |  |  |
| --- | --- | --- | --- |
| **BERT Twitter** | | | |
| *epoch* | *training loss* | *validation loss* | *F1* |
| 1 | 0.627200 | 0.614605 | 0.690475 |
| 2 | 0.461700 | 0.518351 | 0.747963 |
| 3 | 0.322800 | 0.647933 | 0.748337 |
| 4 | 0.201300 | 1.088970 | 0.740506 |
| 5 | 0.118600 | 1.349419 | 0.746362 |

Tabella 11: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche

Figura 11: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Twitter in 5 epoche

Immagine che contiene testo, Diagramma, linea, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Nonostante ciò, i valori di F1-score durante la validazione rimangono relativamente stabili tra 0.74 e 0.75 dalla seconda alla quinta epoca. Questo apparente paradosso è spiegabile con l’iniziale presenza di segnali utili appresi dal training, che si mantengono validi in fase di validazione interna, ma che non si traducono in una reale generalizzazione sul test set ufficiale (*Tabella 10*).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **BERT** | | | | |
|  | *accuracy* | *F1-score (macro avg)* | *support* | *confusion matrix* |
| **CH** | 0.64 | 0.63 | 199 | [[53, 46], [26, 74]] |
| **DI** | **0.69** | **0.68** | 74 | [[20, 17], [6, 31]] |
| **JO** | 0.56 | 0.56 | 200 | [[52, 48], [39, 61]] |
| **TW** | **0.48** | **0.44** | 152 | [[16, 71], [8, 57]] |

Tabella 12: Metriche di valutazione del modello BERT sui dati di test suddivisi per genere

Infatti, le performance sul test evidenziano un notevole crollo: accuracy pari al 48% e F1 macro pari a 0.44, i valori più bassi tra tutti i generi analizzati. La confusion matrix (*Figura 12*) mostra che il modello ha appreso a riconoscere solo una delle due classi (la classe 1), con un recall dell’88% per F, ma una fortissima difficoltà nel riconoscere correttamente la classe 0 (solo 16 esempi su 87).

Questo squilibrio produce una classificazione distorta, in cui la maggioranza delle istanze viene assegnata alla classe F, indipendentemente dalla verità. Il fenomeno può essere legato alla natura rumorosa e non standardizzata del linguaggio usato su Twitter, che rende difficile per BERT individuare segnali stilistici costanti. Inoltre, il test set ridotto (solo 152 esempi) rendere instabili le metriche di valutazione.

Immagine che contiene testo, schermata, Rettangolo, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

Figura 12: Confusion matrix delle performance di BERT sul genere Twitter

**Valutazione complessiva di BERT.** In conclusione, i risultati ottenuti mostrano una forte variabilità delle performance di BERT a seconda del genere testuale, come riportato nella *Tabella 5*.

Il modello raggiunge le prestazioni più elevate su Diary (F1 = 0.68, accuracy = 0.69), seguito da Children (F1 = 0.63), mentre si osserva un calo significativo su Journalism (F1 = 0.56) e un risultato nettamente inferiore su Twitter (F1 = 0.44, accuracy = 0.48). Le confusion matrix rivelano che nei generi Children e Diary BERT riesce a discriminare in modo più bilanciato tra le due classi, con una leggera preferenza per la classe femminile. Al contrario, nei generi Journalism e Twitter, il modello tende a commettere errori sistematici, spesso riconducibili a debolezza dei segnali stilistici marcati (nel caso giornalistico) o a una forte rumorosità linguistica (nel caso di Twitter), che compromette la generalizzazione. In particolare, nel caso di Twitter, l'andamento delle curve di loss e la struttura della confusion matrix suggeriscono un comportamento da overfitting. A ciò si aggiunge la difficoltà strutturale posta dai testi molto brevi tipici di Twitter, che forniscono un contesto linguistico estremamente limitato e potenzialmente insufficiente per un modello come BERT, che si basa sulla ricchezza distribuzionale del testo per inferire caratteristiche latenti. Nel complesso, il comportamento di BERT rispecchia l’eterogeneità dei domini linguistici analizzati: i generi più personali (Diary, Children) offrono segnali più sfruttabili dal modello, mentre quelli più neutri (Journalism) o informali e variabili (Twitter) pongono maggiori sfide alla classificazione del genere.

Conclusioni

Bibliografia

Trevor Cohn, Yulan He, and Yang Liu. 2020. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Association for Computing Machinery. 1893. Association for Computing Machinery. *Computing Reviews*, 24(11):503–512.

Margherita Hack. 2011. Libera scienza in libero Stato. Bur.

Dan Jurafsky and Christopher Manning. 2012. Natural language processing. *Instructor* 2012(988): 3482.

[Figura 1: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Children 2](#_Toc199174721)

[Figura 2: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Diary 3](#_Toc199174722)

[Figura 3: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Journalism 4](#_Toc199174723)

[Figura 4: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Twitter 4](#_Toc199174724)

[Figura 5: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Children in 5 epoche 6](#_Toc199174725)

[Figura 6: Confusion matrix delle performance di BERT sul genere Diary 7](#_Toc199174726)

[Figura 7: Curve di loss del training e della validation di BERT sul genere Diary nelle 5 epoche 7](#_Toc199174727)

[Figura 8: Confusion matrix delle performance di BERT sul genere Diary 8](#_Toc199174728)

[Figura 9: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Journalism in 5 epoche 8](#_Toc199174729)

[Figura 10: Confusion matrix delle performance di BERT sul genere journalism 8](#_Toc199174730)

[Figura 11: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Twitter in 5 epoche 9](#_Toc199174731)

[Figura 12: Confusion matrix delle performance di BERT sul genere Twitter 9](#_Toc199174732)

[Tabella 1: Accuracy per fold dell’SVM rispetto alla baseline dummy su Children (5-fold cross validation) 2](#_Toc199174711)

[Tabella 2: Accuracy per fold dell’SVM rispetto alla baseline dummy su Diary (5-fold cross validation) 3](#_Toc199174712)

[Tabella 3: Accuracy per fold dell’SVM rispetto alla baseline dummy su Journalism (5-fold cross validation) 4](#_Toc199174713)

[Tabella 4: Accuracy per fold dell'SVM rispetto alla baseline dummy su Twitter (5-fold cross validation) 4](#_Toc199174714)

[Tabella 5: Report finale sul test set di SVM ProfilingUD 5](#_Toc199174715)

[Tabella 6: Risultati del fine-tuning di BERT sul genere Children per 5 epoche 6](#_Toc199174716)

[Tabella 7: Risultati del fine-tuning di BERT sul genere Diary per 5 epoche 7](#_Toc199174717)

[Tabella 8: Risultati del fine-tuning di BERT sul genere journalism per 5 epoche 8](#_Toc199174718)

[Tabella 9: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche 9](#_Toc199174719)

[Tabella 10: Metriche di valutazione del modello BERT sui dati di test suddivisi per genere 9](#_Toc199174720)