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.

1 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.

2 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**.

2.1 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.

2.2 Organizzazione dei file(?)

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

2.3 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*

SVM ProfilingUD				
	accuracy	*F1-score (macro avg)*	*support*	*confusion matrix*
CH	0.5729	0.57	199	[[69, 31], [54, 45]]
DI				
JO				
TW				

SVM ProfilingUD				
	accuracy	*F1-score (macro avg)*	*support*	*confusion matrix*
CH	0.5729	0.57	199	[[69, 31], [54, 45]]
DI				
JO				
TW				

3 Modelli e risultati

3.1 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: Accuracy per fold dell'SVM rispetto alla baseline dummy su Children (5-fold cross validation).*Tabella 1*, le performance ottenute risultano altamente instabili tra i fold, con accuracy variabili tra 0.425 e 0.625, mentre i 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 (

), 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 (

) 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

SVM ProfilingUD				
	accuracy	*F1-score (macro avg)*	*support*	*confusion matrix*
CH	0.5729	0.57	199	[[69, 31], [54, 45]]
DI				
JO				
TW				

Tabella 2: Report finale sul test set di SVM ProfilingUD

(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 (***Errore. L'origine riferimento non è stata trovata.***) 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*.

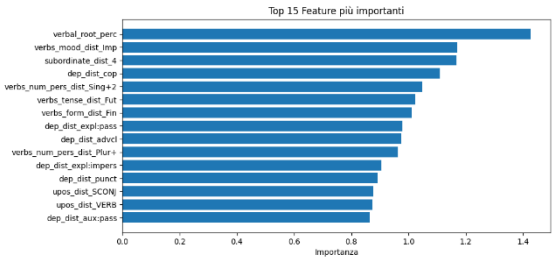

Figura 1: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Children

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.

3.2 SVM con word embeddings

3.3 SVM con n-grammi

LA TABELLA E LA DIDASCALIA 2 SI MUOVONO QUANDO NASCONO NUOVE RIGHE
Forse ho risolto!

3.4 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 3* e visualizzato in *Figura 2*.

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 3: Risultati del fine-tuning di BERT sul genere children per 5 epoche

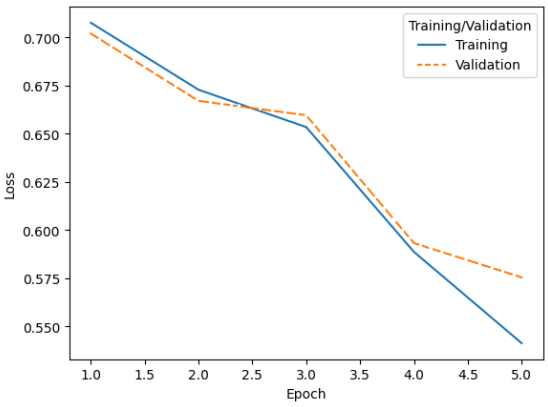

Figura 2: Curve di loss di training e validation durante il fine-tuning di BERT sul genere children in 5 epoche

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 (***Errore. L'origine riferimento non è stata trovata.***) 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.

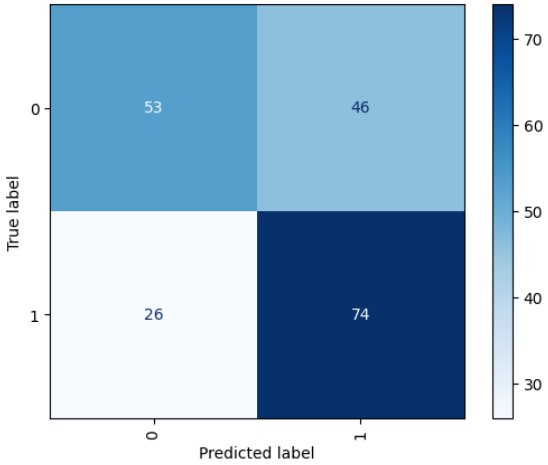

Figura 3: Confusion matrix delle performance di BERT sul genere Children

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 4* e Figura 4.

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

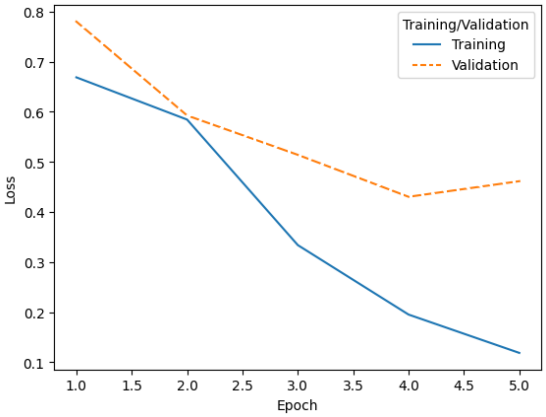

Figura 4: Curve di loss del training e della validation di BERT sul genere Children nelle 5 epoche

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 5*) 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.

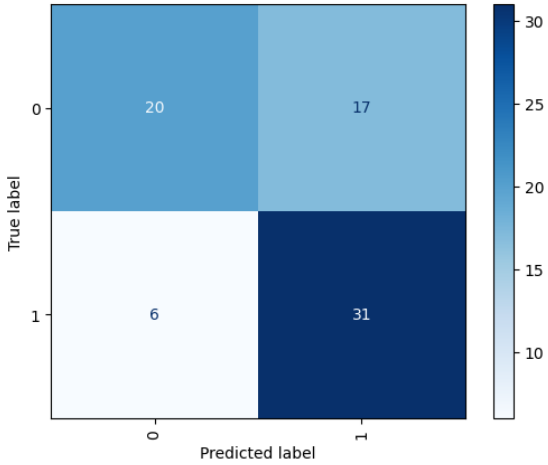

Figura 5: Confusion matrix delle performance di BERT sul genere diary

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.

Journal. 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 5* e *Figura 6*, la validation loss non segue una traiettoria chiaramente decrescente, suggerendo una fase di apprendimento meno stabile rispetto agli altri generi.

BERT Journal			
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 5: Risultati del fine-tuning di BERT sul genere journal per 5 epoche

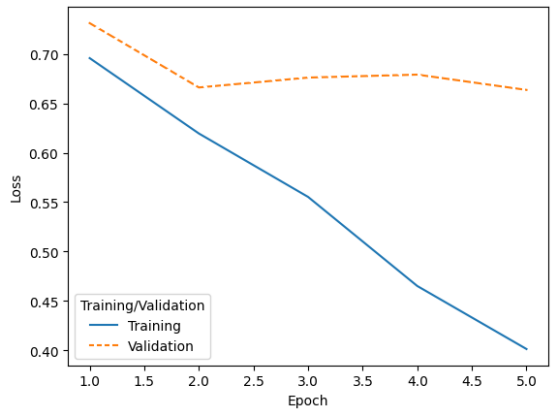

Figura 6: Curve di loss di training e validation durante il fine-tuning di BERT sul genere journal in 5 epoche

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 7*) 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.

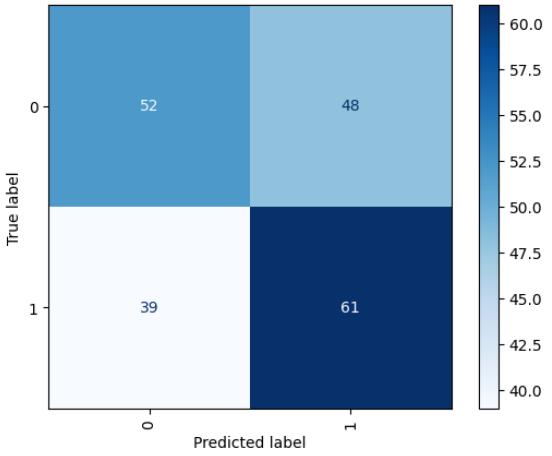

Figura 7: Confusion matrix delle performance di BERT sul genere journal

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 6* e *Figura 8*Figura 7, 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 6: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche

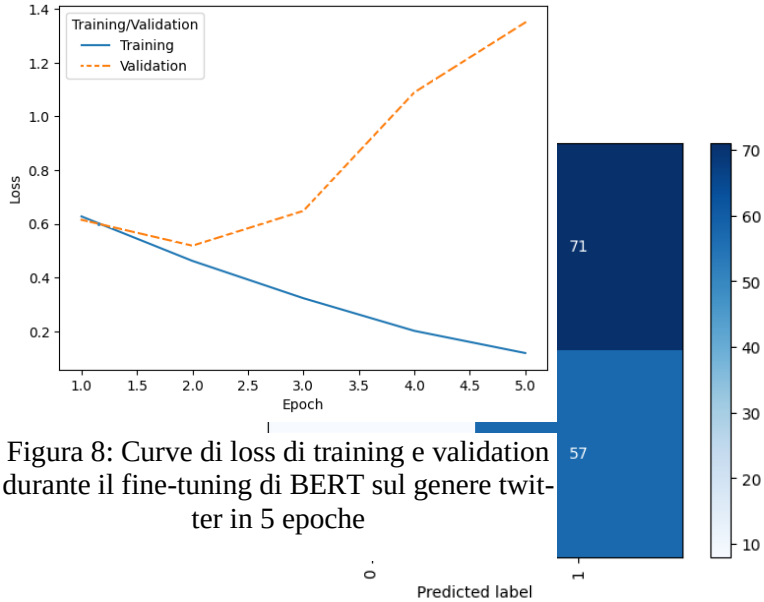

Figura 8: Curve di loss di training e validation durante il fine-tuning di BERT sul genere twitter in 5 epoche

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 7*).

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 7: 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 9*) 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.

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.

Figura 9: Confusion matrix delle performance di BERT sul genere journal

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 sul genere Children..... 2
Figura 2: Curve di loss di training e validation durante il fine-tuning di BERT sul genere children in 5 epoche 3
Figura 3: Confusion matrix delle performance di BERT sul genere Children 4
Figura 4: Curve di loss del training e della validation di BERT sul genere Children nelle 5 epoche..... 4
Figura 5: Confusion matrix delle performance di BERT sul genere diary 5
Figura 6: Curve di loss di training e validation durante il fine-tuning di BERT sul genere journal in 5 epoche 5
Figura 7: Confusion matrix delle performance di BERT sul genere journal 5
Figura 8: Curve di loss di training e validation durante il fine-tuning di BERT sul genere twitter in 5 epoche 6
Figura 9: Confusion matrix delle performance di BERT sul genere journal 6

Tabella 1: Accuracy per fold dell'SVM rispetto alla baseline dummy su Children (5-fold cross validation)..... 2
Tabella 2: Report finale sul test set di SVM ProfilingUD..... 3
Tabella 3: Risultati del fine-tuning di BERT sul genere children per 5 epoche..... 3
Tabella 4: Risultati del fine-tuning di BERT sul genere Children per 5 epoche..... 4
Tabella 5: Risultati del fine-tuning di BERT sul genere journal per 5 epoche..... 5
Tabella 6: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche 6
Tabella 7: Metriche di valutazione del modello BERT sui dati di test suddivisi per genere 6