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. This project addresses the Gender Cross-Genre (GxG) task proposed at EVALITA 2018, which focuses on author profiling in Italian across four textual genres: Children, Diary, Journalism, and Twitter. The aim is to explore the effectiveness of different classification models (SVM and BERT) and representations (morphosyntactic features, n-grams, word embeddings) in predicting the author's gender. Results confirm the difficulty of the task: although most models outperform the 50% baseline in validation, generalization to the official test set remains challenging. Best accuracies vary significantly by genre, rarely exceeding 70%. The textual genres differ in style, lexical complexity, and class distribution, making the task more difficult and leading to considerable variation in model performance across genres

Italiano. Questo progetto affronta il task Gender Cross-Genre (GxG), introdotto da EVALITA 2018, focalizzato sull'author profiling di testi in italiano appartenenti a quattro generi testuali: Children, Diary, Journalism e Twitter. L'obiettivo è valutare l'efficacia di diversi modelli di classificazione (SVM e BERT) e rappresentazioni del testo (feature morfosintattiche, n-grammi, word embeddings) nel predire il genere dell'autore. I risultati confermano la complessità del task: sebbene la maggior parte dei modelli superi la baseline del 50% in fase di validazione, la generalizzazione sul test set ufficiale resta problematica. Le migliori accuracy variano significativamente tra i generi, superando raramente il 70%. La natura cross-genere del task introduce instabilità dovuta a fattori come l'eterogeneità stilistica, la semplicità lessicale o lo sbilanciamento delle classi.

1 Introduction

Negli ultimi anni, la crescente diffusione delle piattaforme digitali ha profondamente trasformato le dinamiche di produzione e diffusione dei contenuti testuali. La possibilità di pubblicare online in modo immediato e pressoché gratuito ha reso l'accesso alla scrittura pubblica estremamente capillare, consentendo a chiunque di condividere testi senza vincoli tecnici o economici rilevanti. Questa democratizzazione dell'accesso alla parola scritta, tuttavia, ha reso più complessa l'attribuzione dell'autorialità: se da un lato è diventato semplice immettere contenuti nel flusso comunicativo digitale, dall'altro identificare l'autore di un testo resta un compito non banale, particolarmente in ambienti caratterizzati da informalità, anonimato o pseudonimia. L'author profiling nasce proprio con questo obiettivo: inferire automaticamente alcune caratteristiche latenti dell'autore (come età, genere o stile comunicativo) a partire esclusivamente dal testo. Tra queste, il genere rappresenta una delle più studiate, e viene comunemente trattato come una variabile binaria (maschio/femmina). In ambito NLP, la predizione del genere ha raggiunto buoni risultati in scenari controllati, soprattutto sui social media, dove il lessico informale e personale tende a far emergere segnali stilistici distintivi. Tuttavia, anche in scenari in-genre (quando l'addestramento e la valutazione avvengono all'interno dello stesso genere testuale) la predizione del genere si rivela tutt'altro che banale. I segnali linguistici associati al genere possono essere deboli, ambigui o distribuiti in modo non uniforme all'interno del corpus, e le performance dei modelli possono variare sensibilmente da un genere all'altro. Analizzare i risultati in-genre consente di valutare con maggiore precisione quali rappresentazioni e modelli riescano a

catturare tratti distintivi tra autori maschili e femminili, tenendo conto delle specificità linguistiche e stilistiche di ciascun dominio. Per affrontare questa sfida, è stato introdotto il task Gender Cross-Genre (GxG) nell'ambito della campagna EVALITA 2018. Il compito consiste nel predire il genere dell'autore su testi italiani appartenenti a generi molto diversi tra loro: Children, Diary, Journalism e Twitter. L'obiettivo non è solo testare la robustezza dei modelli in scenari realistici e variabili, ma anche comprendere quali generi favoriscano o ostacolino l'emersione di tratti linguistici legati al genere.

2 Dataset

2.1 Analisi del dataset

Il dataset utilizzato nel progetto è stato fornito nell'ambito della campagna di valutazione GxG (Gender Cross-Genre), e include testi scritti da autori di genere maschile (M) e femminile (F) in quattro generi testuali: *Children*, *Diary*, *Journalism* e *Twitter*. Come visibile nella *Tabella 1*, la distribuzione dei dati è perfettamente bilanciata per genere in ciascuna categoria: ogni genere testuale include esattamente 100 testi scritti da autori F e 100 da autori M, ad eccezione di *Twitter*, che contiene 3000 testi per ciascun genere, per un totale di 6000 esempi.

	F	**M**	**Totale**	**%F**	**%M**
CH	100	100	200	50	50
DI	100	100	200	50	50
JO	100	100	200	50	50
TW	3000	3000	6000	50	50

Tabella 1: Distribuzione dei testi per genere testuale e genere dell'autore

Questa simmetria assicura che le performance dei modelli non siano influenzate da uno sbilanciamento nella distribuzione delle classi. Un'analisi delle caratteristiche strutturali dei testi (*Tabella 2*) rivela differenze significative tra i generi testuali.

	Media parole	**Media caratteri**	**Media frasi**	**Numero testi**
CH	329.54	1903.24	16.86	200
DI	413.88	2469.52	22.10	200
JO	565.85	3549.89	26.71	200
TW	16.90	117.57	1.62	6000

Tabella 2: Lunghezza media dei testi per genere testuale

I testi *Twitter* risultano estremamente brevi, con una media di circa 17 parole per testo e meno di 2 frasi per messaggio. Al contrario, i testi *Journalism* sono i più lunghi, con una media di 566 parole e circa 27 frasi. I generi *Diary* e *Children* si collocano in una posizione intermedia, con rispettivamente 414 e 330 parole per testo. Anche la varianza interna è piuttosto elevata, soprattutto nel caso dei testi *Diary*, che mostrano una forte eterogeneità in termini di lunghezza (varianza > 100.000 parole). Queste differenze strutturali suggeriscono che ciascun genere testuale presenta caratteristiche stilistiche e comunicative specifiche, che potrebbero influenzare in modo significativo la predizione del genere dell'autore. In particolare, la brevità e informalità di *Twitter* potrebbero limitare l'efficacia di approcci lessicali, mentre testi più lunghi e articolati come quelli *Journalism* o *Diary* potrebbero offrire indizi più robusti, anche a livello sintattico o morfologico.

2.2 Organizzazione dei file

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*

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 3*, le performance ottenute risultano altamente instabili tra i fold, con accuracy variabili tra 0.43 e 0.63, mentre le baseline dummy oscillano tra 0.40 e 0.48. In alcuni casi, il modello supera la baseline dummy, ma senza garantire una generalizzazione robusta.

SVM ProfilingUD Children

fold	*accuracy*	*Dummy classifier*
1	0.63	0.48
2	0.43	0.43
3	0.45	0.48
4	0.53	0.40
5	0.50	0.43

Tabella 3: 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 7*), ottiene un'accuracy di 0.57 e un F1-score macro di 0.57, superando la baseline dummy (0.50) ma restando su valori moderati. La confusion matrix (*Tabella 7*) 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.*

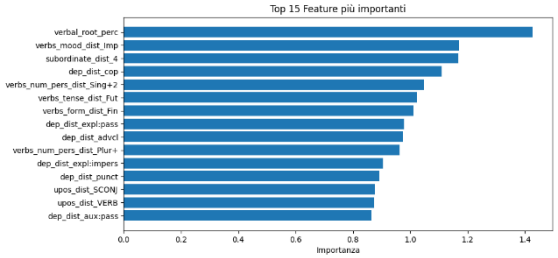

Figura 1: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Children

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 4*, le performance del modello risultano moderatamente stabili tra i fold, con accuracy comprese tra 0.43 e 0.60, mentre le baseline dummy oscillano tra 0.38 e 0.50. 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 4: Accuracy per fold dell'SVM rispetto alla baseline dummy su Diary (5-fold cross validation)

Quando il modello viene valutato sul test set (*Tabella 7)* ottiene un'accuracy di 0.55 e un F1-score macro di 0.55, superando la baseline (0.50) e mantenendo un livello di performance costante con la cross-validation. La confusion matrix (*Tabella 7*) 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 2*) 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-*

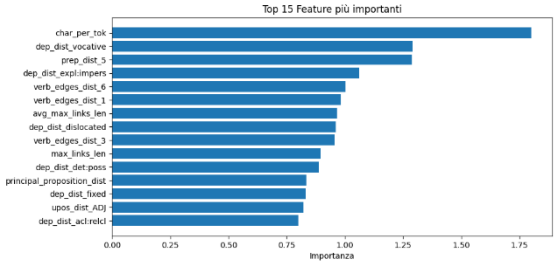

Figura 2: Classifica delle 15 feature più rilevanti per il modello SVM ProfilingUD sul genere Diary

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 5*, 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 0.43 e 0.48. 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	0.68	0.45
2	0.70	0.43
3	0.68	0.48
4	0.5	0.43
5	0.68	0.43

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

Quando valutato sul test set (*Tabella 7*) 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 7*) 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*.

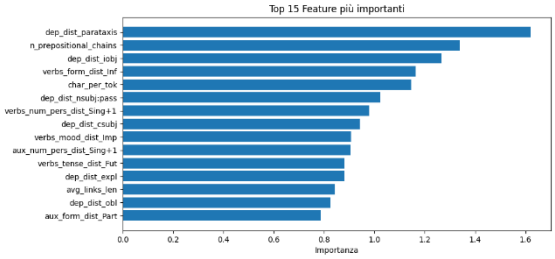

Figura 3: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Journalism

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 7*) 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 6*, le performance del modello SVM risultano relativamente stabili tra i fold, con accuracy comprese tra 0.61 e 0.65, mentre le baseline dummy oscillano tra 0.47 e 0.50. 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	0.63	0.49
2	0.65	0.50
3	0.61	0.50
4	0.63	0.50
5	0.62	0.47

Tabella 6: Accuracy per fold dell'SVM rispetto alla baseline dummy su Twitter (5-fold cross validation)

Quando valutato sul test set ufficiale (*Tabella 7*), 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 7*) 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 (*Figura 4*) 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*.

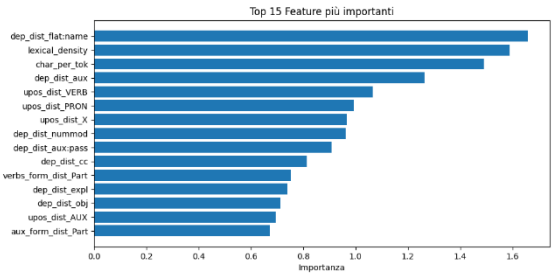

Figura 4: Classifica delle 15 feature più rilevanti per il modello SVM sul genere Twitter

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	0.57	0.57	199	[[69, 31], [54, 45]]
DI	0.55	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 7: 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.

3.2 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 8*, la strategia *mean_noun_adj* si è distinta come la più efficace in fase di cross validation, con una accuracy media di 0.62, mostrando buona robustezza anche tra i fold (0.55, 0.60, 0.70, 0.65, 0.60). 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_noun_adj	0.62
max_all	0.61
mean_all	0.56
mean_verb	0.56
max_noun_adj	0.55
max_verb	0.49

Tabella 8: 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*. Come visibile in *Tabella 12*, l'accuracy ottenuta è stata pari a 0.5427, inferiore a quanto osservato in validazione. Sebbene leggermente superiore alla baseline, il risultato indica una capacità discriminativa solo modesta. La confusion matrix (*Tabella 12*) mostra un numero elevato di errori, con 49 testi di genere F e 42 di genere M classificati erroneamente. Il modello riesce a distinguere parzialmente le due classi, ma la presenza di quasi 100 errori complessivi su 199 testi suggerisce che la rappresentazione adottata non è sufficientemente informativa per generalizzare su dati non visti.

Diary. Come visibile nella *Tabella 9*, la strategia *mean_all* si è distinta come la più efficace in fase di cross-validation, con una accuracy media di 0.620, 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 Diary	
strategia	*mean accuracy*
mean_all	0.62
mean_noun_adj	0.61
max_all	0.61
max_noun_adj	0.58
mean_verb	0.56
max_verb	0.50

Tabella 9: Mean accuracy della cross validation di ciascuna strategia (genere Diary)

Come da specifiche del progetto, sul test set è stata testata esclusivamente la strategia migliore, ovvero *mean_all*. Come visibile in *Tabella 9*, l'accuracy ottenuta è stata pari a 0.70, superiore a quanto osservato in validazione. Il risultato si discosta significativamente dalla tendenza osservata negli altri generi, risultando ben al di sopra della baseline (0.50) e suggerendo una buona generalizzazione del modello su dati non visti. La confusion matrix (*Tabella 12*) mostra una distribuzione degli errori meno problematica rispetto ad altri casi, con 20 testi F e 32 testi M correttamente classificati. Tuttavia, si osservano ancora 17 falsi negativi e 5 falsi positivi. Il modello mostra quindi una maggiore efficacia nel riconoscere la classe maschile rispetto a quella femminile, come confermato dal f1-score (0.74 per M contro 0.65 per F).

Journalism. Come visibile nella *Tabella 10*, la strategia *max_noun_adj* si è distinta come la più efficace in fase di cross-validation, con una accuracy media di 0.63. Le performance ottenute nei singoli fold (0.80, 0.50, 0.63, 0.68, 0.55) risultano però piuttosto instabili. La strategia meno efficace è risultata *mean_verb*, con una media di 0.54.

SVM word embeddings Journalism	
strategia	*mean accuracy*
max_noun_adj	0.63
max_all	0.61
mean_noun_adj	0.60
mean_all	0.57
mean_verb	0.54
max_verb	0.54

Tabella 10: Mean accuracy della cross validation di ciascuna strategia (genere Journalism)

Come da specifiche del progetto, sul test set è stata testata esclusivamente la strategia migliore, ovvero *max_noun_adj*. Come visibile in *Tabella 12*, l'accuracy ottenuta è stata pari a 0.50, inferiore sia alla media di validazione sia alla baseline del dummy classifier (0.50). Questo risultato indica che il modello non è stato in grado di apprendere pattern utili per distinguere in modo efficace tra i due generi, restituendo una performance sostanzialmente casuale. La confusion matrix (*Tabella 12*) conferma la difficoltà del classificatore: 55 testi maschili e 46 femminili sono stati classificati erroneamente. L'elevato numero di errori (101 su 200) e la simmetria quasi perfetta tra le due classi suggeriscono che la rappresentazione adottata - basata su max pooling e filtri grammaticali - non fornisce una struttura discriminativa sufficiente per il genere Journalism, forse a causa della maggiore omogeneità stilistica o della complessità testuale di questa categoria.

Twitter. Come visibile nella *Tabella 11*, la strategia *mean_noun_adj* è stata la più efficace in fase di cross-validation, con una accuracy media di 0.60. Le performance ottenute nei singoli fold (0.60, 0.61, 0.53, 0.63, 0.63) risultano piuttosto consistenti, suggerendo una buona capacità del modello di adattarsi alle diverse suddivisioni del training set. La strategia meno efficace è risultata invece *max_verb*, con una media di 0.48.

SVM word embeddings Twitter	
strategia	*mean accuracy*
mean_noun_adj	0.60
max_noun_adj	0.59
mean_all	0.59
max_all	0.56
mean_verb	0.52
max_verb	0.48

Tabella 11: Mean accuracy della cross validation di ciascuna strategia (genere Twitter)

Sul test set la strategia *mean_noun_adj*, come visibile nella *Tabella 12*, ha ottenuto un'accuracy pari a 0.46, inferiore alla media osservata in validazione, ma leggermente superiore alla baseline del dummy classifier (0.43). Nonostante il lieve vantaggio rispetto alla baseline, il risultato evidenzia una capacità discriminativa piuttosto limitata. La confusion matrix (*Tabella 12*) mostra 57 testi maschili e 25 testi femminili classificati erroneamente, per un totale di 82 errori su 152. Il modello mostra una maggiore tendenza a identificare correttamente i testi femminili (recall = 0.62), ma a scapito della precisione per la classe maschile. La natura estremamente breve dei testi su Twitter, unita al numero ridotto di esempi nel test set (molto inferiore rispetto al training), potrebbe

aver influito negativamente sulle capacità di generalizzazione del modello. La scarsità di segnali linguistici complessi, tipica di questa piattaforma, rende difficile per una rappresentazione basata su word embeddings non contestuali cogliere differenze robuste tra i generi.

Valutazione complessiva dell'SVM con word embeddings. Il modello SVM lineare basato su word embeddings ha mostrato prestazioni eterogenee nei diversi generi testati, con risultati che oscillano tra una buona accuratezza (Diary: 0.70) e performance inferiori alla baseline (Journalism: 0.50). La scelta della strategia ottimale sulla base della cross validation non ha garantito risultati consistenti sul test set, a dimostrazione del fatto che, pur utile, non è sempre predittiva della capacità di generalizzazione su dati realmente nuovi. Le difficoltà maggiori si sono osservate nei generi *Twitter* e *Journalism*, dove le caratteristiche del dominio (brevità estrema dei testi nel primo caso, complessità sintattica e stile informativo nel secondo) sembrano ridurre l'efficacia di una rappresentazione non contestuale come quella offerta da word embeddings statici. Inoltre, la presenza di un numero molto ridotto di esempi nel test set di *Twitter* ha probabilmente accentuato la fragilità del modello, rendendolo più sensibile a errori marginali. In sintesi, il classificatore SVM con word embeddings appare sensibile al genere testuale e alla qualità della rappresentazione adottata. Sebbene possa raggiungere buoni livelli di accuratezza in domini più strutturati o lessicalmente ricchi (*Diary*), risulta meno efficace in contesti dove il segnale linguistico è debole, rumoroso o estremamente compresso. Questi risultati suggeriscono la necessità di esplorare rappresentazioni più robuste e contestuali (come gli embeddings da modelli Transformer) per affrontare generi meno "espliciti" dal punto di vista stilistico e morfosintattico.

SVM word embeddings					
	strategy	*accuracy*	*F1-score (macro-avg)*	*support*	*confusion matrix*
CH	*mean_noun_adj*	0.54	0.54	199	[[51, 49], [42, 57]]
DI	*mean_all*	0.70	0.70	74	[[20, 17], [5, 32]]
JO	*max_noun_adj*	0.50	0.49	200	[[54, 46], [55, 45]]
TW	*mean_noun_adj*	0.46	0.46	152	[[40, 25], [57, 30]]

Tabella 12: Risultati dell'SVM con word embeddings sul test set delle migliori strategie per ciascun genere

3.3 SVM con n-grammi

In questa fase del lavoro è stato implementato un classificatore lineare basato su SVM che prende in input una rappresentazione del testo costruita attraverso l'uso di n-grammi. L'obiettivo è esplorare l'impatto di diverse configurazioni di rappresentazione del testo, intese come combinazioni tra tipo di informazione e lunghezza degli n-grammi, nel distinguere il genere dell'autore. Una volta individuata la rappresentazione migliore in base alla media delle accuracy ottenute in cross-validation, questa è stata testata sul test set ufficiale. Anche in questo caso, i risultati sono stati confrontati anche con una baseline semplice, rappresentata da un dummy classifier che predice sempre la classe più frequente.

Children. Come visibile nella *Tabella 13*, la strategia che usa *trigrammi di POS* si è distinta come la più efficace in fase di cross-validation, con una accuracy media di 0.580. Le performance nei singoli fold (0.43, 0.55, 0.58, 0.60, 0.75) risultano però piuttosto variabili, suggerendo una certa instabilità del sistema, potenzialmente legata a una scarsa coerenza nei pattern grammaticali associati al genere in questo dominio. La strategia meno efficace è risultata l'uso *di trigrammi di lemmi*, con una accuracy media di 0.46.

SVM n-grams Children		
tipo	*n*	*mean accuracy*
POS	3	0.58
POS	1	0.57
lemma	2	0.57
POS	3	0.54
char	3	0.55
word	1	0.51
word	3	0.51
lemma	1	0.51
word	2	0.49
char	2	0.48
lemma	3	0.46

Tabella 13: Accuracy media della 5-fold cross-validation per ciascuna configurazione di n-grammi (genere Children)

Come da specifiche del progetto, sul test set è stata testata esclusivamente la strategia migliore. Come visibile in *Tabella 17*, l'accuracy ottenuta è stata

pari a 0.48, inferiore a quanto osservato in validazione e anche leggermente al di sotto della baseline del dummy classifier (0.50). Questo risultato indica che il modello non è riuscito a generalizzare adeguatamente, nonostante le performance incoraggianti su parte del training set. La confusion matrix (*Tabella 17*) mostra una distribuzione quasi simmetrica degli errori: 52 testi F e 52 testi M sono stati classificati erroneamente, a fronte di 48 e 47 correttamente predetti, rispettivamente. Questo equilibrio segnala una classificazione in gran parte aleatoria, con una capacità discriminativa minima. Le difficoltà del modello possono essere attribuite alla natura lessicalmente semplice dei testi del genere *Children* e alla bassa densità informativa delle sequenze di PoS, che risultano poco indicative rispetto al genere dell'autore.

Diary. Come visibile nella *Tabella 14*, la strategia più efficace in fase di cross-validation è risultata l'uso di *trigrammi di caratteri*, con una accuracy media di 0.83 date le prestazioni molto elevate e stabili tra i fold (0.88, 0.75, 0.90, 0.80, 0.83). Questo risultato indica che le sequenze sub-lessicali sono particolarmente informative nel genere Diary, probabilmente grazie alla varietà morfologica e ortografica che caratterizza questo genere testuale. La strategia meno efficace è risultata l'uso di *trigrammi di lemmi*, con una media pari a 0.60.

SVM n-grams Diary		
tipo	*n*	*mean accuracy*
char	3	0.83
word	1	0.79
lemma	1	0.75
lemma	2	0.73
char	1	0.72
word	2	0.67
char	2	0.66
POS	1	0.64
POS	2	0.64
POS	3	0.63
lemma	3	0.62
word	3	0.60

Tabella 14: Accuracy media della 5-fold cross-validation per ciascuna configurazione di n-grammi (genere Diary)

Sul test set, come visibile in *Tabella 14*, l'accuracy ottenuta è stata pari a 0.60, un valore significativamente inferiore a quello osservato in validazione, ma comunque superiore alla baseline del dummy classifier (0.50). Il calo può essere attribuito principalmente alla ridotta dimensione del test set, che nel caso di Diary è limitata a 74 testi totali, rendendo più instabili le valutazioni. La confusion matrix (*Tabella 17*) mostra uno sbilanciamento evidente nelle predizioni: 21 testi di genere F sono stati classificati erroneamente come M, mentre solo 9 testi M sono stati confusi con F. Il modello mostra quindi una chiara preferenza per la classe maschile, raggiungendo un recall del 76% per M, ma solo del 43% per F. Questo squilibrio potrebbe essere legato a uno sbilanciamento implicito nelle forme utilizzate nei testi o a una maggiore omogeneità stilistica dei testi maschili. Nonostante il buon potenziale mostrato in validazione, il comportamento sul test suggerisce che il modello fatica a mantenere una buona generalizzazione fuori dal training set.

Journalism. Come visibile nella *Tabella 14*, la configurazione più efficace in fase di cross-validation è risultata quella basata su *unigrammi di lemmi*, con una accuracy media di 0.720. I risultati nei singoli fold (0.65, 0.78, 0.75, 0.78, 0.65) sono elevati e abbastanza stabili, suggerendo che i lemmi catturano efficacemente il contenuto informativo dei testi giornalistici, riducendo la variabilità dovuta alla flessione morfologica. La strategia meno efficace è invece risultata quella basata su *bigrammi di POS*, con una accuracy media di 0.53.

SVM n-grams Journalism		
tipo	*n*	*mean accuracy*
lemma	1	0.72
word	1	0.69
word	2	0.69
char	2	0.67
char	3	0.67
lemma	2	0.65
char	1	0.63
POS	1	0.63
POS	3	0.62
word	3	0.57
lemma	3	0.54
POS	2	0.53

Tabella 15: Accuracy media della 5-fold cross-validation per ciascuna configurazione di n-grammi (genere Journalism)

I risultati ottenuti sul test set, riportati in *Tabella 17*, evidenziano un calo dell'accuracy a 0.48, inferiore alla baseline del dummy classifier (0.50). Il modello mostra difficoltà nel generalizzare ai

dati del test set, classificando correttamente solo 55 F e 41 M. Sebbene le prestazioni in validazione fossero promettenti, la generalizzazione si è rivelata modesta.

Twitter. Come sintetizzato nella *Tabella 16*, la configurazione più performante in cross-validation è risultata l'uso di *unigrammi di parole*, con un'accuracy media pari a 0.69. Le cinque accuracy ottenute nei singoli fold (0.70, 0.71, 0.67, 0.68, 0.69) mostrano una variabilità contenuta, suggerendo una discreta stabilità del modello in validazione.

SVM n-grams Twitter		
tipo	*n*	*mean accuracy*
word	1	0.69
char	3	0.68
lemma	1	0.67
char	2	0.65
lemma	2	0.64
word	2	0.63
char	1	0.63
POS	2	0.60
word	3	0.60
POS	1	0.58
lemma	3	0.58
POS	3	0.58

Tabella 16: Accuracy media della 5-fold cross-validation per ciascuna configurazione di n-grammi (genere Twitter)

Come da specifiche del progetto, sul test set è stata valutata esclusivamente la strategia migliore. L'accuracy ottenuta è pari a 0.42, nettamente inferiore sia alla media di cross-validation sia alla baseline del dummy classifier (0.57). La confusion matrix (*Tabella 17*) evidenzia un forte squilibrio delle predizioni: 61 testi F sono stati classificati correttamente a fronte di soli 3 testi M, mentre 84 testi M vengono erroneamente assegnati al genere F. Questo pattern rivela che il classificatore ha imparato caratteristiche lessicali fortemente associate al genere F negli esempi di training e tende a etichettare come femminili anche la maggior parte dei testi maschili. Gli unigrammi di parola imparano dettagli troppo specifici dei tweet di training che poi raramente si ritrovano nei tweet di test. Poiché i tweet sono brevi e molto vari nello stile, questi indizi lessicali da soli non bastano a distinguere il genere.

SVM n-grams

	tipo	*n*	*accuracy*	*F1-score (macro avg)*	*support*	*confusion matrix*
CH	POS	3	0.48	0.48	199	[[48, 52], [52, 47]]
DI	char	3	0.60	0.5	74	[[16, 21], [9, 28]]
JO	lemma	1	0.48	0.48	200	[[55, 45], [59, 41]]
TW	word	1	0.42	0.32	152	[[61, 4], [84, 3]]

Tabella 17: Risultati dell'SVM con n-grams sul test set delle migliori strategie per ciascun genere

Valutazione complessiva dell'SVM basato su n-grammi. Nel complesso, gli SVM basati su n-grammi mostrano buone potenzialità in fase di validazione ma limiti evidenti di generalizzazione. Ogni genere premia un diverso tipo di rappresentazione, ma passando dal training al test set quasi tutti i modelli subiscono un calo netto di accuratezza e, salvo il caso del genere *Diary*, non riescono a superare in modo significativo la baseline del dummy classifier. Questo indica che gli n-grammi catturano pattern fortemente dipendenti dal dominio, difficilmente estendibili a testi nuovi, sia per l'eterogeneità stilistica sia per la semplicità lessicale o per lo sbilanciamento delle classi. In sintesi, gli SVM con n-grammi si confermano un buon punto di partenza per l'analisi di genere, ma per ottenere prestazioni stabili su domini diversi occorre usare informazioni più ricche, come caratteristiche stilistiche più complesse, rappresentazioni del testo basate sul contesto o metadati, e adottare metodi migliori per evitare l'overfitting.

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 18* e visualizzato in *Figura 5*.

BERT Children			
epoch	*training loss*	*validation loss*	*F1*
1	0.71	0.70	0.33
2	0.67	0.67	0.65
3	0.65	0.66	0.67
4	0.59	0.59	0.75
5	0.54	0.58	0.75

Tabella 18: Risultati del fine-tuning di BERT sul genere Children per 5 epoche

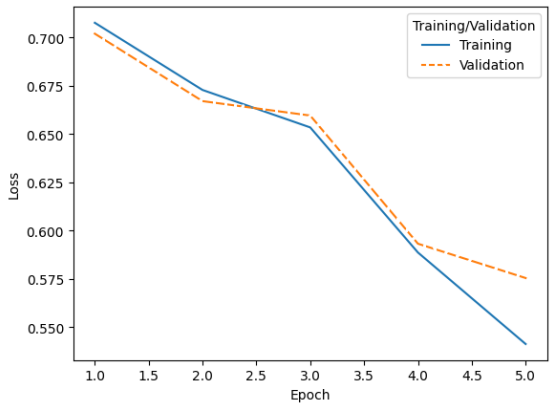

Figura 5: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Children in 5 epoche

In particolare, si osserva una discesa regolare della validation loss da 0.70 a 0.58, con un andamento parallelo al training loss, che raggiunge un minimo di 0.54 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 (*Tabella 22*) 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. Questo sbilanciamento potrebbe riflettere delle caratteristiche stilistiche più marcate o riconoscibili nei testi femminili, portando il 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 19* e *Figura 6*.

BERT Diary			
epoch	*training loss*	*validation loss*	*F1*
1	0.67	0.78	0.17
2	0.58	0.59	0.73
3	0.33	0.51	0.78
4	0.20	0.43	0.83
5	0.12	0.46	0.83

Tabella 19: Risultati del fine-tuning di BERT sul genere Diary per 5 epoche

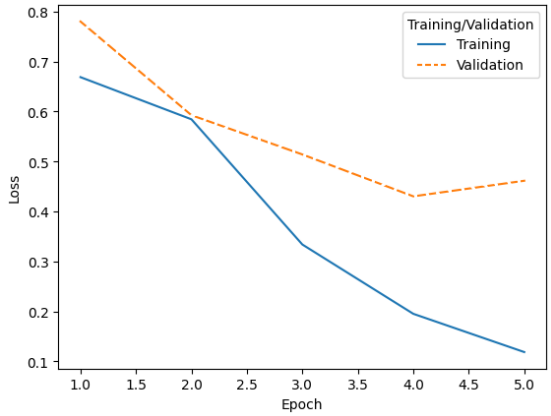

Figura 6: Curve di loss del training e della validation di BERT sul genere Diary nelle 5 epoche

La training loss cala in modo consistente da 0.67 a 0.12, mentre la validation loss registra una riduzione iniziale da 0.78 a 0.43 alla quarta epoca, seguita da un lieve peggioramento nell'ultima epoca (0.46), 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 0.83 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 (*Tabella 22*) 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. 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 0.70 a 0.40. La validation loss si stabilizza tra valori compresi tra 0.66 e 0.68 dopo un iniziale calo nella seconda epoca. Come mostrato in *Tabella 20* e *Figura 7*, 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	0.70	0.73	0.22
2	0.62	0.67	0.60
3	0.56	0.68	0.56
4	0.47	0.68	0.62
5	0.40	0.66	0.65

Tabella 20: Risultati del fine-tuning di BERT sul genere journalism per 5 epoche

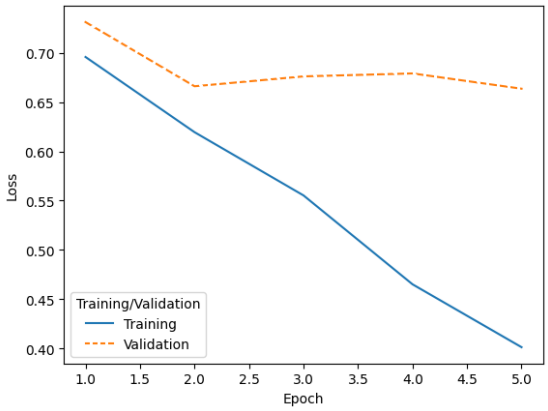

Figura 7: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Journalism in 5 epoche

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 (*Tabella 22*) 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. 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 0.63 a 0.12, mentre la validation loss, dopo un iniziale miglioramento fino alla seconda epoca (0.52), inizia a peggiorare rapidamente, raggiungendo valori molto alti (1.35 alla quinta epoca). Come mostrato in *Tabella 21* e *Figura 8*, 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	0.63	0.61	0.69
2	0.46	0.52	0.75
3	0.32	0.65	0.75
4	0.20	1.09	0.74
5	0.12	1.35	0.75

Tabella 21: Risultati del fine-tuning di BERT sul genere Twitter per 5 epoche

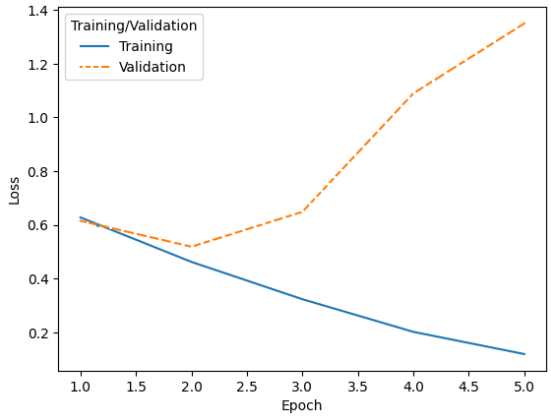

Figura 8: Curve di loss di training e validation durante il fine-tuning di BERT sul genere Twitter in 5 epoche

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 22*).

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 22: 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 (*Tabella 22*) 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.

Conclusioni

I risultati ottenuti confermano la complessità del task di gender prediction, anche in contesto in-genre, e mostrano come le prestazioni dei modelli siano fortemente condizionate dal genere testuale. I modelli testati – SVM con feature morfosintattiche, n-grammi, word embeddings e BERT – ottengono risultati migliori nei generi più lunghi e strutturati, come Diary e Children, mentre incontrano maggiori difficoltà in contesti brevi e stilisticamente instabili come Twitter, dove si osserva una marcata tendenza a classificare i testi come femminili. Anche Journalism, nonostante la maggiore lunghezza dei testi, mostra una discriminabilità limitata, probabilmente per effetto di uno stile più neutro e contenuti meno soggettivi. Le feature morfosintattiche di Profiling-UD forniscono una base linguistica interessante, ma si rivelano limitate se utilizzate isolate. Gli n-grammi mostrano buone performance ma scarsa robustezza fuori campione. L'approccio con word embeddings evidenzia la necessità di rappresentazioni che vadano oltre la forma delle parole e siano in grado di catturare aspetti legati al significato, al contesto d'uso e alle scelte stilistiche. Modelli più avanzati come BERT rispondono parzialmente a questa esigenza, ottenendo le migliori performance in validazione, pur mostrando una certa vulnerabilità all'overfitting nei testi più brevi. Nel complesso, la predizione del genere resta un compito sfidante che richiede una combinazione di segnali lessicali, sintattici e contestuali, adattati alle specificità del dominio testuale. Approcci futuri dovranno integrare segnali linguistici più ricchi e strategie di regolarizzazione più robuste per migliorare la generalizzazione dei modelli.