Sviluppo di modelli di classificazione per affrontare il compito di authorship attribution

Denise Atzori

Corso di Linguistica computazionale II

Sessione estiva 2025

Introduzione

Questa relazione illustra lo svolgimento di un compito di authorship attribution su un corpus creato a partire dalle opere di narrativa di tre autrici italiane di fine ottocento:

- Anna Maria Zuccari, meglio conosciuta come Neera (1846-1918)
- Matilde Serao (1856-1927)
- Maria Antonietta Torriani, meglio conosciuta come Marchesa Colombi (1840-1920)

Le autrici sono state scelte sia per l'importanza delle loro opere nel panorama letterario italiano, sia per la vicinanza di stili e temi trattati, che rende il compito di authorship attribution più complesso e di potenziale interesse letterario[1].

Il compito è stato svolto usando librerie Python e in quattro fasi:

1. Addestramento di un classificatore basato su Support Vector Machine (SVM) lineare e informazioni linguistiche non lessicali
2. Addestramento di un classificatore basato su SVM lineare e n-grammi
3. Addestramento di un classificatore basato su SVM lineare e word embedding
4. Fine-tuning di un modello di linguaggio neurale (Bert-base-italian-cased)

A partire dalla costruzione del dataset, la relazione illustra ogni fase di svolgimento del compito, soffermandosi sui risultati e sulle deduzioni ricavate.

[1] Le tre autrici hanno partecipato attivamente al clima di fermento politico e letterario di fine ottocento, spesso entrando in contatto tra loro e condividendo temi e battaglie, come quella per l'emancipazione femminile. Si veda a riguardo, ad esempio: (Mitchell, 2008)

1. Costruzione del dataset

Le opere del dataset sono state scaricate in file txt e codifica UTF-8 dal Progetto Gutenberg. Di ciascuna autrice è stato scelto un numero variabile di opere, in funzione della disponibilità del sito e della lunghezza. Per ogni autrice sono stati selezionati 1000 paragrafi per il training set, 100 per il validation set e 100 per il test set. Per garantire l'uniformità dei testi, sono state scelte solo opere in prosa. Di seguito l'elenco delle opere selezionate per ciascuna autrice:

- **Matilde Serao**: *Piccole anime* (1883), *La conquista di Roma* (1885), *Gli amanti* (1894), *Le amanti* (1894), *L'infedele* (1897)
- **Marchesa Colombi**: *Tempesta e bonaccia: romanzo senza eroi* (1877), *La cartella n.4* (1880), *Senz'amore* (1883), *Cara Speranza* (1888), *I ragazzi d'una volta e i ragazzi di adesso* (1888), *Serate d'inverno* (1917)
- **Neera**: *Nel sogno* (1893), *L'amuleto* (1897), *La vecchia casa* (1900), *Le idee di una donna* (1904), *Rogo d'amore* (1914)

1.1 Pulizia dei testi

Prima di procedere all'estrazione dei paragrafi, i testi sono stati normalizzati attraverso uno script per eliminare i ritorni a capo singoli interni a ogni paragrafo; sono inoltre stati eliminati i ritorni a capo superiori ai doppi per facilitare le successive fasi di estrazione.

Ogni testo è stato poi sottoposto a un trattamento manuale con l'ausilio di regex per eliminare paratesti (introduzioni e conclusioni inserite dal Progetto Gutemberg, informazioni di collana etc) e delimitatori di aree testuali non utili ai fini dell'analisi.[2]

1.2 Estrazione dei paragrafi e creazione dei file

Il dataset è stato creato selezionando paragrafi di lunghezza compresa tra 50 e 100 token, approssimando i token usando lo spazio come divisore. Data la scarsità del materiale a disposizione sul Progetto Gutemberg (massimo sei opere per autrice, e di lunghezza spesso esigua), si è rivelato necessario spezzare al primo punto disponibile i paragrafi più lunghi di 100 token, per estrarne sotto porzioni adatte all'analisi.

[2] Un esempio è dato dai tre asterischi che in molti casi spezzano a metà un capitolo per dargli respiro. Data la natura del tutto arbitraria di questo tipo di marcatori, si è preferito procedere a un controllo visivo e non automatico dei testi.

Al fine di assicurare la corretta valutazione dei modelli, per ciascuna autrice è stata scelta un'opera dedicata esclusivamente al validation e al test set[3], e non presente dunque nel file di training. Ogni paragrafo è stato salvato su un file txt con il nome della sua autrice, un id e il set di cui è parte.

2. Classificatore basato SVM lineari e informazioni linguistiche non lessicali

Per l'addestramento del primo classificatore si è fatto ricorso alle informazioni linguistiche non lessicali estratte dal tool Profiling UD (Brunato et all, 2020), che ha fornito i file dei paragrafi annotati in formato CONNLU e il profilo linguistico, composto da 139 features per ciascun paragrafo, in differenti livelli di descrizione linguistica.

Le features estratte dal tool sono state normalizzate al fine di renderle confrontabili tra loro e poi sottoposte a un Support Vector Classifier Lineare in forma di vettori di lunghezza 139 features.

2.1 Valutazione del modello

Per valutare le performance del modello si è prima eseguita una 5-fold cross-validation sul dataset di addestramento (training set). Le performance di ogni iterazione sono state confrontate con una baseline ottenuta tramite un DummyClassifier con strategia "most-frequent" (accuracy 0.32). In ogni iterazione, il modello ha ottenuto un'accuracy dello 0.70 circa.

La valutazione delle performance del modello sul validation e sul test set è stata condotta attraverso un classification_report e una matrice di confusione per ciascuno. Sul validation set, il modello ha ottenuto un'accuracy dello 0.68. Sul test dello 0.65.

Dalle matrici di confusione si può notare che il modello tende, in entrambi i casi, a riconoscere con sicurezza leggermente maggiore i testi di Serao, mentre confonde gli altri attribuendoli alternativamente alle tre autrici; una possibile ragione potrebbe essere stilometrica, ovvero una maggiore distanza dello stile di Serao dallo stile delle altre due autrici.

[3] Le opere selezionate per i set di valutazione sono "L'amuleto" per Neera, "La conquista di Roma" per Serao e "Tempesta e Bonaccia" per Colombi. Le tre opere sono state scelte in ragione della loro lunghezza.

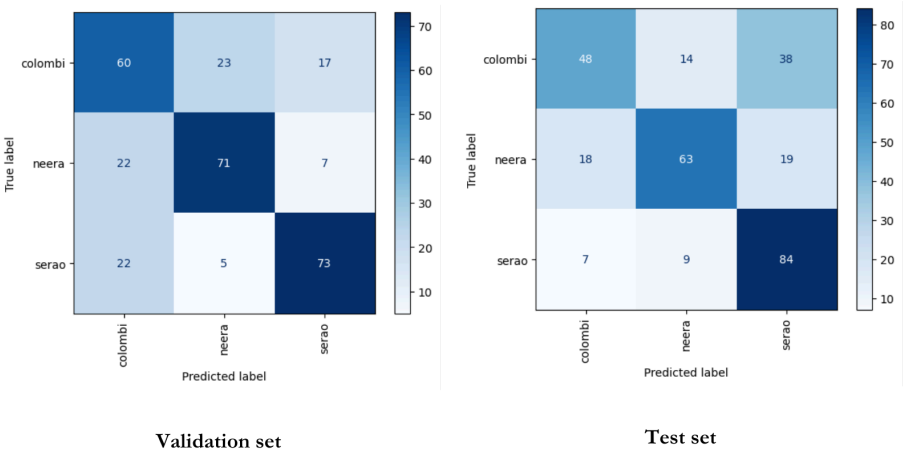

Figura 1: Matrici di confusione della SVC addestrata con informazioni linguistiche non lessicali

Per ciascuna delle tre autrici sono state stampate le venti features più rilevanti, per osservare quali informazioni linguistiche sono state utilizzate dal modello per il compito di classificazione:

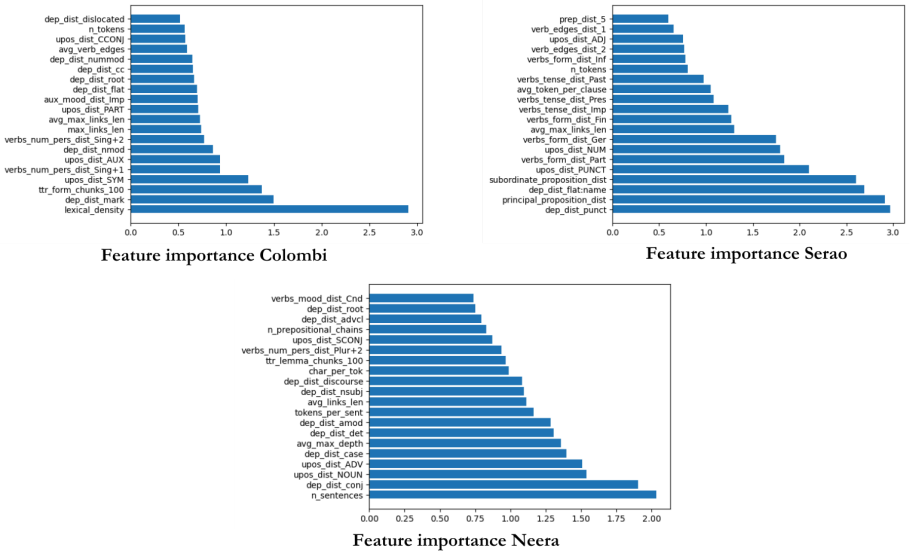

Figura 2: Feature importance per le tre classi - SVC addestrata con informazioni linguistiche non lessicali

Dalla Figura 2 si può osservare che il modello usa come feature più rilevante per riconoscere Colombi l'informazione morfo-sintattica della *lexical density*, ovvero il rapporto tra parole contenuto su tutte le parole presenti; per riconoscere Neera hanno grande rilevanza il numero di frasi, ovvero una "Raw Text Property", e la *dep_dist_conj*, un'informazione di relazione sintattica che rappresenta la distribuzione media della relazione che coinvolge le congiunzioni; infine, per

Serao hanno grande rilevanza, tra le altre molto vicine, la *dep_dist_punct,* anch'essa relazione sintattica che misura la distribuzione media delle relazioni che coinvolgono la punteggiatura e la *principal_proposition_dist*, una misura dell'uso della subordinazione che indica la distribuzione delle frasi principali.

3. Classificatore basato su SVM lineari e n-grammi

Il secondo classificatore è stato progettato sempre a partire da un SVC, al quale sono state somministrate rappresentazioni dei testi basate su n-grammi di caratteri, parole e part-of-speech, estratte dai file CONNLU creati nel task precedente.

I vettori di features sono stati normalizzati sulla frequenza, per tenere conto della diversa lunghezza dei documenti, ed è stato applicato un filtro per eliminare le features poco frequenti.

Sono state testate diverse configurazioni di features, al fine di valutare sul validation set la più performante:

- 1 e 2-grammi di parole e 1-2 grammi di caratteri (accuratezza dello 0.80)
- 1 e 2-grammi di parole e 1 e 2-grammi di lemmi (accuratezza dello 0.81)
- 1 e 2-grammi di parole e 1 e 2-grammi di POS (accuratezza dello 0.85)

3.1 Valutazione del modello

La valutazione sul test set è stata condotta sulla configurazione più performante, ovvero 1 e 2-grammi di parole e 1 e 2-grammi di POS, ottenendo un'accuratezza dello 0.86.

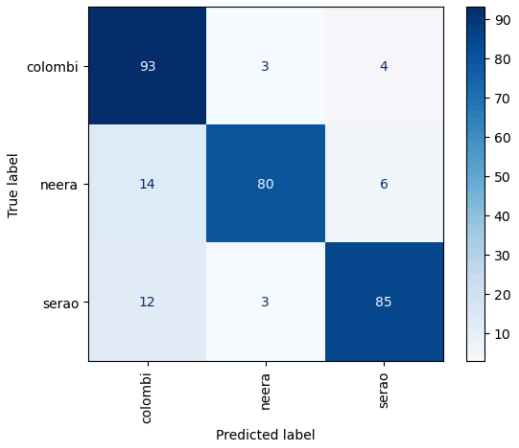

Figura 3: Matrice di confusione della SVC sul test set addestrata con vettori di 1 e 2-grammi di parole e 1 e 2-grammi di POS

In questo caso, la matrice di confusione mostra che il modello ha identificato con buona sicurezza i testi di Colombi, facendo invece più errori su Serao e Neera; l'analisi delle features più rilevanti (Figura 4) offre buoni spunti di riflessione: per riconoscere Colombi, ad esempio, il modello si è basato sulla presenza di una congiunzione preceduta da punteggiatura (WORD_2_,_e) e (POS_2_PUNCT_CCONJ), che potrebbe rappresentare un'effettiva cifra stilistica dell'autrice; nel caso di Neera, viene considerata rilevante l'alta presenza dell'elisione dell'articolo determinativo maschile (WORD_1_l'), insieme all'uso di alcuni nomi propri ricorrenti e di altre forme di punteggiatura (WORD_1_? e WORD_1_!); per Serao, infine, l'uso dei due punti (WORD_1_:) e delle virgole (WORD_1_,) sembrano essere caratteristiche di distinzione rispetto alle altre due autrici.

È bene comunque segnalare che, trattandosi di testi estratti da un progetto non accademico e non controllato, è possibile che alcune di queste caratteristiche dipendano dalla trascrizione dei testi effettuata e vadano dunque soppesate e raffrontate ai testi originali per accertarne la validità.

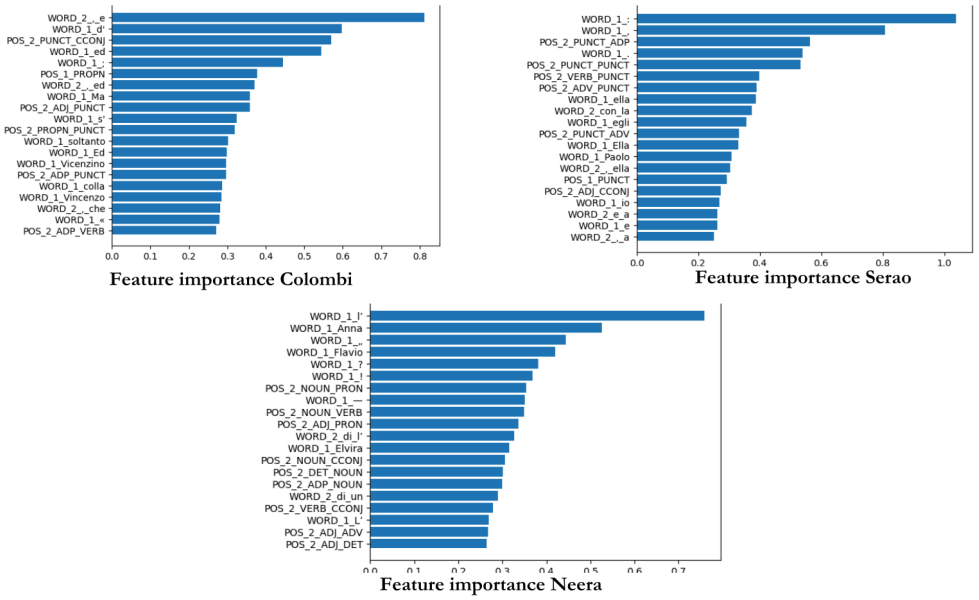

Figura 4: Feature importance per le tre classi - SVC addestrata con vettori di 1 e 2-grammi di parole e 1 e 2-grammi di POS

4. Classificatore basato su SVM lineari e word embedding

La terza analisi è stata svolta su un Support Vector Classifier al quale sono state sottoposte rappresentazioni vettoriali dei testi costruite attraverso word embeddings. Le rappresentazioni sono

state scaricate dal progetto "Italian Word Embeddings" dell'ItaliaNLP Lab e si è scelto di utilizzare i word embeddings da 128 dimensioni.

Sono state sperimentate diverse strategie di aggregazione degli embeddings di ogni paragrafo, per trovare la più performante sul validation test:

- Media di tutti gli embeddings di ogni paragrafo (accuratezza dello 0.70)
- Medie degli embeddings di aggettivi, nomi e verbi concatenati (accuratezza dello 0.43)
- Medie dei word embeddings di aggettivi, nomi, verbi, ausiliari, preposizioni e pronomi, concatenati (accuratezza dello 0.50)
- Somma dei word embeddings di ogni paragrafo (accuratezza dello 0.68)
- Massimo tra word embeddings di ogni paragrafo (accuratezza dello 0.43)

4.1 Valutazione del modello

Le prestazioni sul test set sono state calcolate con classification report e matrice di confusione sulla configurazione più performante, ovvero la media semplice di tutti gli embeddings di ogni paragrafo, ottenendo un'accuratezza dello 0.69. In Figura 5 la matrice di confusione ottenuta, che mostra maggiore incertezza nel riconoscimento dei testi di Neera:

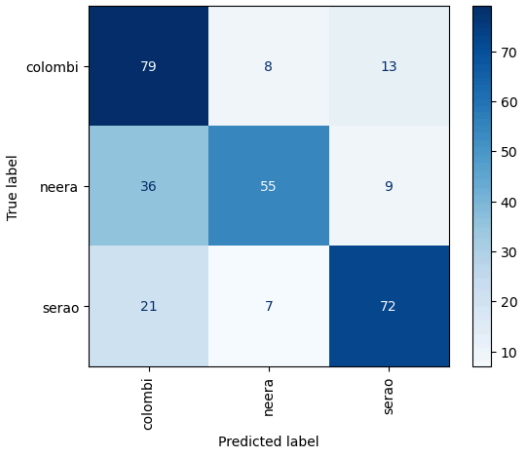

Figura 5: Matrice di confusione della SVC sul test set addestrata con le medie semplici degli embeddings

I risultati delle prestazioni delle diverse configurazioni fanno ipotizzare che, ai fini del riconoscimento dell'autorialità di queste tre autrici, l'SVC si sia basato su informazioni veicolate sia dalle parole piene che da quelle funzionali; l'esperimento condotto usando solo le parole piene

mostra infatti che, con tutta probabilità, non sono queste a veicolare le informazioni che permettono la disambiguazione. Ha senso considerare che, trattandosi di testi ottocenteschi, l'alta presenza di forme d'uso desuete potrebbe essere alla radice della minore informatività degli embeddings creati usando solo le parole piene: è infatti possibile che alcune delle parole caratteristiche usate dalle autrici non siano presenti nelle rappresentazioni vettoriali utilizzate, che sono state create a partire da corpora moderni e non del dominio letterario in analisi.

5. Fine-tuning di un modello di linguaggio neurale (Neural Language Model)

Per l'ultimo task è stato selezionato il modello Bert base italian cased, sviluppato dal MDZ Digital Library team (dbmdz) della Bavarian State Library. Il modello, composto da 111 milioni di parametri, è stato addestrato su un corpus composto da pagine Wikipedia e dall'OPUS corpora, un corpus parallelo open source.

I testi in esame, in formato dataset, sono stati tokenizzati e sottoposti al modello in sequenze di massimo 512 token (subwords). Il modello di Bert è stato fine-tunato per sei epoche e le sue prestazioni sono state calcolate sul validation set con le metriche F-score e osservando le variazioni della Training e Validation Loss lungo le epoche (riportate nella Tabella 1).

3.1 Valutazione del modello

Epoch	Training Loss	Validation Loss	F1
1	0.500900	1.181714	0.688177
2	0.141400	1.512529	0.682161
3	0.042000	1.904724	0.665320
4	0.009900	1.525543	0.735531
5	0.001400	1.760435	0.715480
6	0.000300	1.713900	0.733776

Tabella 1: Performance di Bert base italian cased sul dataset lungo le sei epoche di fine tuning

Come si può osservare dalla Tabella e dalla Figura 5 qui di seguito, nel corso del fine-tuning l'errore sui dati di addestramento è calato fino ad arrivare quasi a zero, mentre quello sui dati del validation set non è mai sceso, segno che il modello non è riuscito a estrarre informazione certa per disambiguare i testi esaminati.

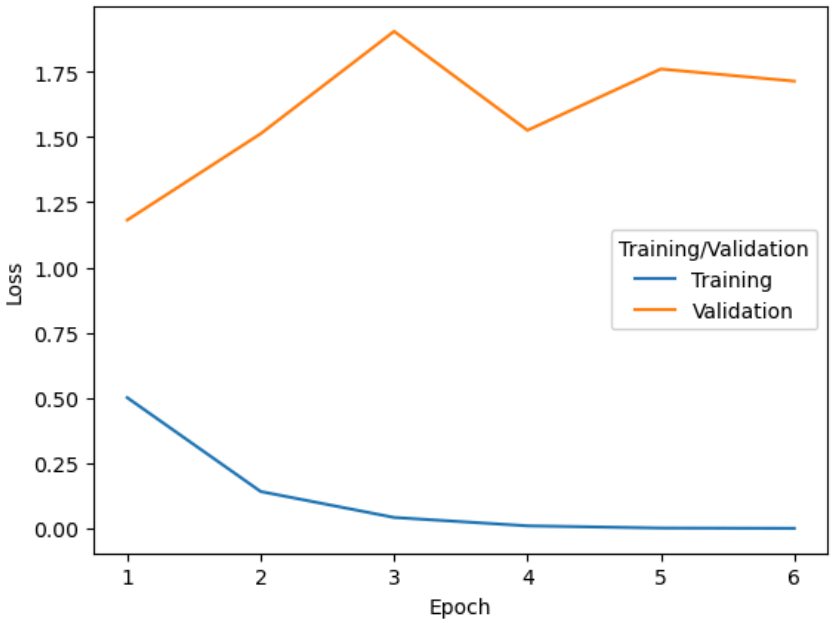

Figura 5: Andamento di Training e Validation Loss di Bert base italian cased lungo le sei epoche di fine tuning

Al termine del fine-tuning è stata valutata l'accuratezza del modello sul test set, che ha ottenuto un valore di 0.69. La matrice di confusione sul test set in Figura 6 mostra che il modello ha riconosciuto correttamente quasi tutti i testi di Colombi, ha commesso alcuni errori su Serao ma ha avuto problemi con Neera, riconoscendo solo 40 dei 100 paragrafi dell'autrice.

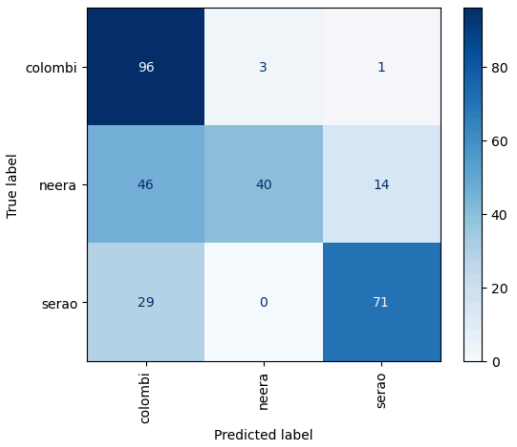

Figura 6: Matrice di confusione di Bert base italian cased sul test

Conclusioni

Come premesso, le tre autrici in esame - Neera, Serao e Colombi - sono state scelte in primo luogo per la vicinanza dei loro temi e delle loro scritture, al fine di rendere il task di classificazione più vicino possibile a un uso reale[4]. Questa scelta ha senz'altro influito sui risultati ottenuti in queste analisi. Dalla vicinanza stilistica tra le autrici, infatti, potrebbero derivare le prestazioni non eccellenti dei modelli, che tendono a confondere gli stili e si mantengono a prestazioni non elevate; fatta eccezione per il modello a n-grammi, che si comporta meglio degli altri in linea con i risultati descritti in letteratura, ad esempio da Daelemans (Daelemans, 2013) che riporta la loro capacità di offrire "an excellent tradeoff between sparseness and information content".

Osservando le matrici di confusione degli ultimi due task, si può notare che i modelli addestrati con features implicite tendono a confondere maggiormente gli scritti di Colombi e Neera, mentre le produzioni di Serao sembrano dare qualche certezza in più. Un'ipotesi di questo comportamento potrebbe essere la maggiore somiglianza tra le scritture delle prime due scrittrici, dovuta all'ambiente culturale e sociale frequentato: mentre Neera e Colombi hanno vissuto entrambe al Nord Italia, muovendosi tra Milano e Torino, Serao ha trascorso la sua vita tra Napoli e Roma, entrando meno in contatto con l'ambiente sociale delle altre due. Per confermare questa ipotesi, uno stadio successivo di analisi potrebbe essere la ricerca nei testi di forme d'uso tipiche dei diversi ambienti sociali, per verificare se davvero sussiste una differenza in tal senso riscontrabile e se può costituire un elemento di distinzione almeno per quanto riguarda Serao.[5]

Nella valutazione delle prestazioni globali, ha inoltre senso notare che l'abbondanza di forme d'uso letterarie ottocentesche nei romanzi esaminati ha di certo aumentato la complessità del task, soprattutto per i modelli basati su features implicite. Il caso del fine-tuning di Bert è in questo senso molto emblematico: come si può notare, infatti, il modello mostra performance ben più basse rispetto ad altri task di classificazione più comuni su corpora moderni (solo 0.69), e non sembra in grado di ricavare informazioni solide sulla base delle quali disambiguare le scritture delle tre autrici.

[4] Una possibile applicazione alla quale si è pensato, in fase di costruzione del dataset, è stata quella della disambiguazione tra le scritture delle tre autrici nel caso di articoli di giornale non firmati ma potenzialmente attribuibili a una delle tre autrici, che si dedicavano anche al giornalismo.

[5] È bene però precisare che, come riporta (Mitchell 2014), le tre autrici scrivevano tutte in un italiano privo di forme dialettali o regionalismi, accessibile anche fuori dal loro contesto sociale e culturale di riferimento.

Allo stesso modo, anche l'SVC addestrato con word-embeddings si ferma a un'accuratezza dello 0.70, cosa che supporta l'ipotesi di un problema legato al vocabolario in uso dalle tre autrici, molto differente da quello conosciuto da Word2vec o dallo stesso Bert in uso.

Possibili strade future potrebbero essere la selezione di modelli neurali addestrati su dataset più vicini a quello in esame; l'implementazione di strategie di adattamento al dominio letterario, con metodi di *Domain Adaptation* semi supervisionato come descritti in (Dell'Orletta, Venturi 2016); e ancora la combinazione, in fase di addestramento dei modelli di machine learning, di informazioni implicite ed esplicite per creare vettori di features più informativi.

I risultati ottenuti nell'addestramento dell'SVC con n-grammi e con features linguistiche non lessicali, infatti, mostrano quanto efficaci siano in questo compito di classificazione caratteristiche stilistiche quali lunghezza delle frasi, densità lessicali, distribuzione media delle congiunzioni, uso della virgola prima delle congiunzioni etc; queste potrebbe offrire una buona base dalla quale partire per un studio futuro sui testi delle tre autrici che combini prospettive computazionali e conoscenze stilometriche e letterarie, allineato agli studi già presenti in letteratura nel campo della *Computational Stylometry*.

Bibliografia

Brunato D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G. (2020) "Profiling-UD: a Tool for Linguistic Profiling of Texts". In *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*, 11-16 May, 2020, Marseille, France.

Daelemans, W. (2013). "Explanation in computational stylometry". In *Computational Linguistics and Intelligent Text Processing*, pages 451–462, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dell'Orletta F., Venturi G. (2016) "ULISSE: una strategia di adattamento al dominio per l'annotazione sintattica automatica". In E. M. Ponti e M. Baudassi (a cura di) *Computer parler soigner: tra linguistica e intelligenza artificiale*, Atti del convegno 15-17 dicembre 2014, Pavia University Press, pp. 55-79.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. 2021. "Universal Dependencies". *Computational Linguistics*, 47 (2): 255–308.

Mitchell, K. (2008). "La Marchesa Colombi, Neera, Matilde Serao: Forging a Female Solidarity in Late Nineteenth-Century Journals for Women". In *Italian Studies*, 63(1), 63–84. https://doi.org/10.1179/007516308X270137

Mitchell, K. (2014). *Italian Women Writers: Gender and Everyday Life in Fiction and Journalism, 1870-1910*. University of Toronto Press.