"The Federalist Papers": analisi stilometrica per l'attribuzione dell'autore tramite tecniche di *machine learning*

Francesca Schiavone
0001145396

2024

1 Introduzione

I *Federalist papers* sono una raccolta di 85 articoli scritti tra il 1787 e il 1788, sotto lo pseudonimo Publius. L'obiettivo di questi articoli era convincere i membri dell'assemblea dello Stato di New York a ratificare la Costituzione degli Stati Uniti d'America. Dietro lo pseudonimo si celavano tre figure chiave: Alexander Hamilton, James Madison e John Jay. Nel 1944, lo storico Douglass Adair condusse delle ricerche per attribuire la paternità ai saggi che venne poi confermata nel 1964. Questo lavoro ha l'obiettivo di attribuire i testi ai rispettivi autori utilizzando tecniche di machine learning basate su caratteristiche stilometriche. La questione centrale è verificare se sia possibile identificare l'autore di un'opera basandosi esclusivamente sull'analisi stilistica. La stilometria, infatti, studia le metriche stilistiche di un testo, come la lunghezza media dei tokens o le parole più utilizzate. Il modello di machine learning impiegato ha raggiunto un'accuratezza del 82%, con una deviazione standard di 0.055 punti percentuali.

2 Metodologia

Caratteristiche stilistiche e Dataset
Le funzioni sviluppate nel file *features.py* si basano su una pipeline preaddestrata della libreria SpaCy, chiamata *en_core_web_trf*. Questo modello,

grazie all'uso di transformer, è stayo scelto per la sua superiore accuratezza ai modelli più piccoli di SpaCy. Il modello linguistico ha svolto diverse operazioni preliminari, tra cui la tokenizzazione. Successivamente, è stato scritto il codice per estrarre le principali caratteristiche stilometriche. Queste caratteristiche includono il conteggio dei tokens e delle frasi nel testo, nonché l'analisi dei tokens e dei bigrammi più frequenti. Nello specifico, sono stati considerati tre approcci: il primo conteggia dei tokens/bigrammi più frequenti (includendo le parole grammaticali, che sono le più comuni in ogni testo). Il secondo passo esclude la punteggiatura dal conteggio della frequenza e infine, il terzo si concentra sull'analisi stilistica escludendo le *stopwords* (parole comuni di un testo, come articoli, verbi ausiliari etc) per mettere in luce le espressioni specifiche degli autori. Si è misurato inoltre il TTR(Type-Token Ratio) che misura la varietà del vocabolario usato da un autore e il rapporto tra il numero di tokens totali e il numero di frasi del singolo testo, che fornisce indicazioni sulla lunghezza dei periodi. Infine, è stata calcolata la frequenza delle categorie grammaticali (Part Of Speech, PoS) che permette di distinguere tra stili più descrittivi, ad esempio caratterizzati da un maggiore uso di aggettivi e stili più narrativi, dove prevalgono verbi.

Queste funzioni sono state applicate a tutti gli 85 testi e utilizzando la libreria *pandas*, è stato creato un dataset con 85 righe e 12 colonne. Ogni colonna contiene una delle metriche descritte tranne l'ultima, denominata 'Target', che invece indica il nome dell'autore del testo. Per l'attribuzione degli autori, ci si è affidati all'elenco fornito da Wikipedia, ignorando eventuali dispute storiche che non rientrano nell'ambito di questo studio.

Modello e valutazione

Prima di applicare il modello di machine learning, è stato essenziale eseguire alcune operazioni preliminari in modo ottimale. La prima operazione svolta è stata il bilanciamento delle classi attraverso l'*upsampling*. Nelle prime fasi si è notato che l'autore 'Jay' era sotto-rappresentato e rischiava di non comparire nel test set. Per risolvere questo problema il numero dei campioni di Jay è stato aumentato casualmente sino a raggiungere il numero della seconda classe maggioritaria, 'Madison'. Successivamente, è stata effettuata la vettorizzazione dei dati non-numerici utilizzando il *CountVectorizer*, che trasforma il testo in una matrice di conteggi dei termini. Questo metodo ha permesso di rappresentare le caratteristiche stilistiche in modo strutturato. I dati numerici sono stati invece standardizzati. Il modello scelto per questa analisi è stato il *Random Forest*. Questo modello ha dimostrato di offrire ottime prestazioni grazie alla sua capacità di gestire diverse tipologie di dati e alla sua robustezza nel ridurre il rischio di *overfitting*; quest'utilimo aumen-

tato in seguito all'*upsampling*. Per valutare il modello, è stata utilizzata la *cross-validation* con *Stratified K-Fold*, che ha permesso di stratificare i dati e ottenere una misura più affidabile delle sue prestazioni. Questo approccio ha prodotto una media dell'accuratezza di circa 0.825 e una deviazione standard di 0.055, confermando la stabilità e l'affidabilità del modello. I risultati della valutazione del modello sono stati i seguenti:

- Accuracy: 0.86,
- Precision, Recall e F1-Score:
- Hamilton: Precisione 0.77, Recall 1.00, F1-Score 0.87
- Jay: Precisione 1.00, Recall 1.00, F1-Score 1.00
- Madison: Precisione 1.00, Recall 0.50, F1-Score 0.67

Le metriche complessive mostrano che il modello ha ottenuto buoni risultati, con una precisione media di 0.90, un recall medio di 0.86 e un F1-Score medio di 0.85.

La matrice di confusione generata mostra la performance del modello ri-

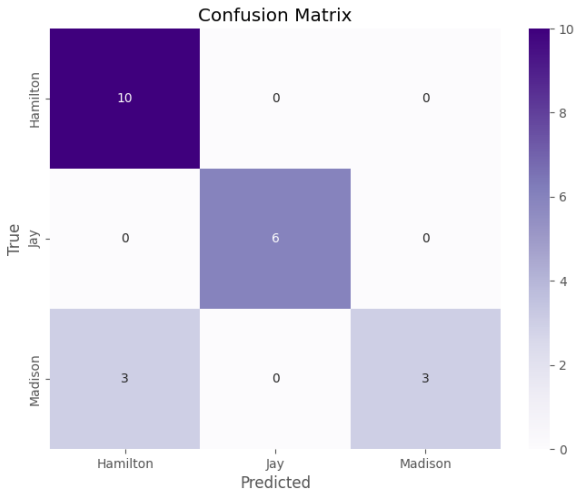

Figura 1: Matrice di confusione del modello Random Forest

spetto a ciasun autore: il modello riconosce senza difficoltà i testi di Jay, ma mostra incertezza tra Hamilton e Madison, i cui stili risultano più simili.

3 Conclusione

Il modello di machine learning ha dimostrato la possibilità di attribuire la paternità di un testo attraverso un'attenta analisi stilometrica dell'autore di riferimento. I risultati ottenuti (un'accuratezza del 82%) mostrano l'efficacia sia del modello selezionato che delle metriche scelte. Tuttavia, ci sono diverse aree in cui questo lavoro potrebbe essere migliorato come nella gestione del possibile *overfitting* e attraverso un più efficiente bilanciamento delle classi con l'utilizzo di tecniche più avanzate. Si potrebbe inoltre confrontare il modello di machine learning con modelli più avanzati per osservare possibili miglioramenti nelle prestazioni.

4 Bibliografia

- *The Federalist Papers*. Wikipedia. URL: `https://en.wikipedia.org/wiki/The_Federalist_Papers`
- W. K. Grieve, "Introduction to Stylometry with Python". Programming Historian, 2021. URL: `https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python`
- Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, 2019.